Bayesian inference of model error for the calibration of two-phase CFD codes

Nicolas Leoni

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX033

Thèse de doctorat

Bayesian inference of model error for the calibration of two-phase CFD codes

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°574 École doctorale de mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées

Thèse présentée et soutenue à Saclay, le 4 avril 2022, par

NICOLAS LEONI

Composition du Jury :

Josselin Garnier Professeur, CMAP, Ecole polytechnique	Président du jury
François Bachoc Maître de conférences, IMT, Université Paul Sabatier	Rapporteur
Youssef Marzouk Professeur, MIT Center for Computational Science and Engineering	Rapporteur
Paola Cinnella Professeur, Sorbonne Université	Examinateur
Pietro Marco Congedo Directeur de recherche, Inria-CMAP, Ecole polytechnique	Directeur de thèse
Olivier Le Maître Directeur de recherche, CNRS-CMAP, Ecole polytechnique	Codirecteur de thèse
Maria-Giovanna Rodio Ingénieur de recherche, Commissariat à l'Energie Atomique	Invitée

Remerciements

Une thèse, c'est bien plus qu'un ensemble de productions scientifiques : c'est un reflet très personnel, presque intime, de trois ans de vie de son auteur. Il m'est impossible, en quelques lignes, de rendre compte de l'ampleur de la reconnaissance que je porte à chacune des personnes qui ont, directement ou indirectement, contribué à l'aboutissement de ma thèse. Je vais tout de même m'essayer à l'exercice.

En premier lieu je remercie les membres du jury, notamment les rapporteurs : François Bachoc et Youssef Marzouk, qui ont relu ce manuscrit avec attention, l'ont commenté et en ont proposé une critique juste qui a permis d'enrichir son contenu. Merci également à Paola Cinnella et à Josselin Garnier pour l'intérêt qu'ils ont porté à mon travail, illustré par les nombreuses questions pertinentes posées lors de la soutenance.

C'est avec grand plaisir que je remercie Pietro Congedo, mon directeur de thèse, le capitaine du bateau, dont l'optimisme, la persévérance, et la détermination m'ont porté et stimulé durant ces années, avec une bienveillance sans faille. Il a su assurer une direction de proximité tout en créant l'espace nécessaire à mon gain en autonomie, assurant ainsi ma formation scientifique. Mes remerciements vont également à Olivier Le Maître, pour son exigence de rigueur mathématique et méthodologique qu'il a su me transmettre lors de nos nombreuses discussions. Nos innombrables échanges, ponctués d'interrogations, de remises en question, et d'idées salvatrices (en apparence), ont donné à cette thèse la forme qu'elle possède aujourd'hui. Mon trio d'encadrement est complété par Maria-Giovanna Rodio, qui a courageusement représenté la mécanique des fluides dans ce projet, apportant un peu de couleur à un travail soporifiquement mathématique. J'ai grandement bénéficié de son implication constante, et de sa volonté à toujours faire l'effort supplémentaire pour me permettre de travailler dans les meilleures conditions.

Au CEA de Saclay, j'ai été accueilli dans le Service de Thermohydraulique et de Mécanique des Fluides, dont je remercie le chef, Pierre Gavoille, ainsi que l'ensemble de l'équipe, notamment Philippe Fillion et Benoît Guizard, avec qui il fut extrêmement plaisant d'échanger.

Entrons dans le détail avec le Laboratoire de Modélisation et Simulation en Mécanique des Fluides, terreau fertile où les graines de stagiaires germent en jeunes pousses de chercheurs. J'ai eu la chance d'en faire partie et de rencontrer des personnes remarquables qui, par leurs qualités professionnelles et personnelles, ont constitué pour moi des modèles de chercheurs et chercheuses. Commençons par Raksmy, l'homme aux mille talents culinaires, sportifs, artistiques, scientifiques et vestimentaires; Alan l'apôtre d'Emacs et pourfendeur de Vim; et Antoine l'inventeur des bulles montées sur ressorts. Merci également à Élie pour sa joie de vivre contagieuse, à Guillaume et toute sa petite famille, et à Nicolas Dorville, chef du laboratoire, pour avoir toujours laissé sa porte de bureau grande ouverte. Un grand merci à Claire, Erell, Julie, Maria-Adela, Maria-Giovanna (seconde fois) et Stéphanie; ainsi qu'à Alain, André, Jean-Christophe, Jérôme,

Florian, Michel, Nikos, Pascal, Philippe, Pierre-Emmanuel, et Ulrich. Il convient également de remercier les secrétaires, Violette et Katy, qui m'ont toujours aidé avec le sourire dans les quelques démarches administratives que j'ai eu à effectuer.

Au détour d'une pause café nous retrouvions les membres du Laboratoire de Génie Logiciel pour la Simulation, dirigé par Erwan Adam, que je remercie pour sa grande disponibilité. Ce fut un plaisir de pouvoir bénéficier de conseils bayésiens dispensés par Guillaume, Jean-Marc et Rudy. Merci à Clarisse, Christophe, JB et Christian pour les bons moments passés ensemble. Dédicace à Geoffrey, jeune arrivé mais pourtant déjà connu au bataillon.

En parlant de bataillon, le STMF compte dans ses rangs de vaillants doctorants qui bûchent dur pour faire avancer la science et obtenir le titre de docteur à placer dans leur signature mail. Je pense notamment à l'infatigable Aurore, Aymeric le vil *Vim*ien, Werner le roi du *SHMUP*, Antonin le fondu*e* de vélo, Andrew et son *Hot Chocolate Effect*, Téo le tête-en-l'air d'Avignon, Gabriel le ponctuel, Mathis le requin, et Clément le digne successeur. À l'occasion du séminaire des doctorants nous avons festoyé au *Food Asia* auprès de Riccardo, Cassiano, Arthur, Antoine, Oumar et Chufa. Une pensée aux jeunes arrivés qui débutent l'aventure, Jiayi, Nathalie, Loïc, Sébastien, Corentin, Alexis, Vincent, Tanguy et Pierre-Louis. Je vous souhaite à tous de profiter un maximum, de vous donner à fond, et j'espère que mon exemple vous rassurera en observant qu'à la fin, on l'obtient cette thèse, même lorsque l'autorisation de soutenance tombe le jour même à 8h17 pour une présentation à 14h.

J'ai pu croiser au LMSF de nombreux stagiaires et post-doctorants, qui contribuent très activement à la vie hors labo. Merci à la *génération poulailler*: Axelle, Loria, Annamisia, Clément, Marcellin et Rémi, ainsi que Pauline, Estanislau, Étienne, Stéphane, François et Mahran. Je remercie également Songzhi, Harshit, Vinod, Ravik et Evangelos.

De l'autre côté du plateau de Saclay, j'ai eu l'occasion de travailler au Centre de Mathématiques Appliquées, initialement au sein de l'équipe DEFI, devenue IDEFIX, dont je remercie les membres. Merci particulièrement au chef, Houssem Haddar, ainsi qu'à Marcella, Xiaoli, Jing-Rebecca, et Lucas pour leur accueil chaleureux. Je remercie l'ensemble de l'équipe du CMAP, spécialement Nasséra pour son aide précieuse dans les formalités administratives, ainsi que les professeurs qui m'ont permis de travailler en tant que tuteur dans leurs cours.

Moving on to an international context, I wish to thank all members of the newly-formed PLATON team at Inria Saclay, with whom I had great pleasure to work. Thank you Anabel (and JB) for your Bayesian companionship. I sincerely wish, João, that I had tried the famed *francesinha* with you while you were in Paris. I'll come to Portugal soon! Thank you Giulio for the funny moments we had in the office. I thank the Spanish duo, Barbara and Christian, as well.

Un merci français à mes prédécesseurs que j'ai eu la chance de rencontrer au détour d'une conférence ou d'un voyage, François, Nassim et Mickaël, ainsi qu'à mes compagnons de route : Paul, Malo et Élie. Merci à ceux que j'ai eu l'occasion de connaître lors de nos

visioconférences d'équipe, notamment Lalaina qui retrouve avec moi les joies du travail en présentiel. Je dois également beaucoup à Hanadi pour avoir sauvé *in extremis* ma soutenance.

J'ai eu la chance d'être toujours soutenu par mes amis, présents dans l'ensemble des moments difficiles que peut comporter une thèse. Trop nombreux pour tous les citer, potes de prépa, copains d'école, camarades de master, anciens colocs, compagnons de randonnée, votre soutien a énormément compté pour moi. Je mentionnerai brièvement Alixe la férue de *Mario Kart*; monsieur Montiel et sa famille, à qui je souhaite bonne chance pour sa thèse; Sandrine et les danseuses & danseurs du CAO pour les spectacles inoubliables; et la bande de la mission Opéra, Alix, Anouk, GG, Antoine, Guigui, Coco, et aussi FB, qui, au mépris des règles élémentaires de probabilité, pioche toujours la même carte de rôle au *Time Bomb*.

Au cours de ces années de thèse, et même depuis bien avant, j'ai pu compter sur ma famille pour m'encourager : mes grand-mères et mon grand-père, mes oncles et tantes d'Arles, de Nîmes, de Tignes ou des Pennes-Mirabeau, je vous en remercie. Une pensée à mes cousins et cousines, et bon courage à celle qui a l'audace de se lancer dans la recherche elle aussi. Merci également à la famille Jacolin pour leur hospitalité et leur bienveillance. J'ai bénéficié du soutien sans réserve de ma soeur lors de nos nombreuses parties de jeu, de ma mère lors de nos cours de yoga ou à distance depuis le Moyen-Orient, et bien sûr de mon père, qui a consenti à me laisser poursuivre une thèse en statistiques, à condition quand même d'y faire figurer un brin de mécaflu.

Terminons par une petite histoire : dans les premiers mois de ma thèse, alors que j'écoutais attentivement un cours de mathématiques donné à l'occasion d'une école d'été, la porte de l'amphithéâtre s'ouvrit soudainement pour laisser place à une étudiante retardataire, qui vint en trombe s'asseoir à côté de moi. Mon regard se posa sur elle et il ne l'a plus quittée depuis. Merci Sophie d'être à mes côtés, ma plus grande supportrice, ma guide de rando, ma goûteuse de champagne, mon 太阳. Je t'aime et je suis impatient de découvrir les nouvelles aventures que nous allons vivre ensemble.

Publications

Journal papers

- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (2022). "Bayesian Calibration with Adaptive Model Discrepancy". (submitted).
- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (2022). "Efficient Calibration of a CFD code with Adaptive Model Dicrepancy". (submitted).

Conferences

- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (14/01/2021). "Bayesian Inference of Model Error in Imprecise Models". WCCM-ECCOMAS 2021, Paris (online), hal-03119715.
- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (07/05/2019). "Bayesian Calibration of a Two-Group IATE Model for Air-Water Flows". Code_Saturne and NEPTUNE_CFD user meeting, Saclay.

Seminar

- N. Leoni, M.G. Rodio, P.M. Congedo and O. Le Maître (09/2021). "Evaluation de l'erreur de modèle: application à un essai DEBORA". Neptune_CFD technical meeting, Saclay.
- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (01/2021). "Bayesian Inference of Model Error", CMAP PhD students seminar, Palaiseau.
- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (10/2020). "Bayesian Inference of Model Error for CFD Codes", CEA PhD students seminar, Saclay.
- N. Leoni, P.M. Congedo, O. Le Maître and M.G. Rodio (06/2020). "Calibration of CFD Codes with Model Error", EDMH PhD Days, Saclay.

Poster

- N. Leoni, M.G. Rodio, P.M. Congedo and O. Le Maître (10/2021). "Estimation of model error: DEBORA test case". Neptune_CFD technical meeting, Saclay.

Contents

List of Figures

List of Tables

List of Acronyms

BEPU Best Estimate Plus Uncertainty
BWR Boiling Water Reactor
CEA Commissariat à l'Energie Atomique
CFD Computational Fluid Dynamics
CHF Critical Heat Flux
DNS Direct Numerical Simulation
DoE Design of Experiments
EDF Electricité de France
FD Fixed Domain asymptotics
FMP Full Maximum a Posteriori
fUQ forward Uncertainty Quantification
GP Gaussian Process
hGP Gaussian Process for optimal Hyperparameters
ID Increasing Domain asymptotics
IET Integral Effect Tests
IRSN Institut de Radioprotection et de Sûreté Nucléaire
iUQ inverse Uncertainty Quantification
KDE Kernel Density Estimation
KL Kullback-Leibler
KOH Kennedy-O'Hagan
LES Large Eddy Simulation
LHS Latin Hypercube Sampling
LOOCV Leave-One-Out Cross-Validation
M-H Metropolis-Hastings
MCMC Monte-Carlo Markov Chain
MITB MIT Boiling Model
ML Maximum Likelihood
nDoE numerical Design of Experiments
NIGP Noisy Input Gaussian Process
NVG Net Vapor Generation
OAT One at A Time
ONB Onset of Nucleate Boiling
OSV Onset of Significant Vapor
PCA Principal Component Analysis
PWR Pressurized Water Reactor
QMC Quasi Monte-Carlo
RANS Reynolds-Averaged Navier Stokes
RD Repeated Domain asymptotics
SET Separate Effect Tests
TH Thermal-Hydraulics
UQ Uncertainty Quantification
V&V Validation and Verification
V&V&UQ Validation, Verification, and Uncertainty Quantification

List of Symbols

Operators

$\lvert \cdot \rvert$	Determinant of a matrix
$\mathbb{E}[\cdot]$	Expected value of a random variable
$\mathrm{GP}(\cdot, \cdot)$	Gaussian Process distribution
$IG(\cdot, \cdot)$	Inverse Gamma distribution
$D_{\mathrm{KL}}(\cdot \lvert\lvert \cdot)$	Kullback-Leibler divergence
$\mathrm{N}(\cdot, \cdot)$	Normal distribution, 1D or multivariate
$\mathrm{N}_t(\cdot, \cdot)$	Truncated normal distribution
$\mathrm{p}(\cdot)$	Probability Density Fuction
$\mathrm{p}^*(\cdot)$	Unnormalized Probability Density Function
$\mathrm{Tr}(\cdot)$	Trace of a matrix
$\mathrm{Var}[\cdot]$	Variance of a random variable

Statistical quantities

$\mathbb{B}$	Space of mean parameters
β	Mean of model discrepancy, mean of a Gaussian Process
$\mathbf{C}$ or $\mathbf{V}$ or $\mathbf{\Gamma}$	Covariance matrix of a Gaussian Process or a Multivariate Normal distribution
$c_{\boldsymbol{\psi}}(\cdot, \cdot)$	Covariance function of a Gaussian Process
D	Design of Experiments (set of points)
δ	Dirac function
ε	Measurement error
f	Computer model
$\mathbf{f}_{\boldsymbol{\theta}}$	Vector of evaluations of the computer model
$\boldsymbol{\gamma}$	Random parameter
$\mathbf{H}$	Mean matrix of a Gaussian Process
$\mathbf{h}$	Mean function of a Gaussian Process
$\mathbb{I}$	Background information
$\mathbf{I}_n$	Identity matrix of size n
$\mathbf{K}$	Covariance matrix with added diagonal term
L	Likelihood criterion, likelihood function
l	Correlation length of a Gaussian Process
$\boldsymbol{\mu}$	Mean of a Multivariate Normal distribution
ν	Regularity hyperparameter
ω	Resampling weights
$\mathbf{P}$	Projection matrix
π	Weight of a density mixture
$\boldsymbol{\psi}$	Hyperparameters of model discrepancy, hyperparameters of a Gaussian Process
$\widehat{\boldsymbol{\psi}}$	Estimator of hyperparameters

σ	Standard deviation of model error
σ_{mes}	Standard deviation of measurement error
τ	Self-correlation distance in a Markov Chain
Θ	Space of computer model calibration parameters
$\boldsymbol{\theta}$	Computer model calibration parameters
$\boldsymbol{\theta}^*$	True value of the computer model calibration parameters
$\mathbb{X}$	Space of computer model inputs
$\mathbf{x}$	Computer model inputs
$\mathbf{Y}$	Data matrix
y	True process
y_{obs}	Observed process
$\widetilde{\mathbf{Y}}$	Centered data matrix
z	Model discrepancy
$z_{\boldsymbol{\theta}}$	Adaptive model discrepancy

Physical quantities

a_i	Interfacial area
α	Void fraction
D_{d}	Departure bubble diameter
D_{m}	Mean Sauter bubble diameter
ε	Turbulence dissipation rate
Γ	Heat transfer
Ja_{sup}	Jakob number
ϕ	Heat flux
ρ	Density
T	Temperature
θ_{c}	Static contact angle
V	Velocity (liquid or gas)

Résumé en français

Modèles scientifiques, erreur et inductivisme

Un modèle scientifique est une représentation abstraite du monde qui nous entoure, une traduction mathématique des phénomènes naturels observés à des fins d'interprétation et prédiction. Par définition même, un modèle n'est qu'une approximation et il est impossible de décrire un phénomène naturel avec une précision infinie, en rendant compte avec exactitude de la totalité des interactions physiques qui le composent. Un modèle sera donc toujours entaché d'une erreur d'approximation, ou erreur de modèle, due aux hypothèses simplificatrices sur lesquelles il repose. La présence de cette erreur ne doit pas être une incitation à abandonner la méthode scientifique, au motif que la vérité absolue soit inaccessible : le plus souvent, l'erreur est suffisamment faible pour que les prédictions finales aient un degré de précision suffisant pour l'application souhaitée. Le choix d'utiliser un modèle pour une application donnée requiert de trouver un équilibre entre sa complexité d'utilisation (par exemple, le temps de calcul nécessaire) et le degré de précision permis. Dans certains cas, même le modèle le plus complexe à disposition n'atteint pas la précision souhaitée : on parle alors d'incertitude épistémique, liée au manque de connaissances. La solution est alors de proposer un nouveau modèle plus détaillé.

Une question centrale est de savoir déterminer le domaine de validité d'utilisation d'un modèle scientifique : pour quels phénomènes peut-on avoir confiance en ses prédictions ? En suivant les principes du modèle inductiviste de la science, il n'est jamais possible d'apporter la preuve formelle de validité d'un modèle pour une application donnée, avant d'avoir réalisé cette application. Cependant, il est possible d'avoir un fort niveau de confiance dans le fait que les nouvelles prédictions soient bonnes, par exemple lorsque d'autres applications similaires ont été précédemment réalisées avec succès. Ce niveau de confiance peut être exprimé formellement, et même calculé, en utilisant le langage de l'inférence bayésienne. En empruntant le cadre mathématique des probabilités, l'inférence bayésienne permet de calculer le niveau de confiance dans un modèle, comparer plusieurs modèles entre eux, ou comparer différentes valeurs de paramètres possibles à utiliser dans un modèle. Le théorème de Bayes est un moyen fondamental d'inclure des observations expérimentales pour modifier les degrés de croyances préexistants, et il est au centre de toutes les méthodes statistiques *data-based* qui seront présentées dans cette thèse.

Les méthodes bayésiennes d'estimation d'erreur de modèle présentent en apparence un caractère paradoxal : d'un côté, leur formulation est générale et universelle, et s'applique à toute situation où l'on dispose de données expérimentales et d'un modèle apte à les reproduire, indépendamment du domaine d'application qui peut être la mécanique des fluides comme l'économie ou bien l'épidémiologie. D'un autre côté, l'utilisation de ces méthodes requiert une connaissance poussée du domaine applicatif, car il est nécessaire de savoir formuler des avis sur les paramètres du modèle à faire varier,

la forme générale de l'erreur de modèle, ou encore les autres types d'incertitude présents dans le calcul. Nous illustrons dans cette thèse le principe fondamental que les méthodes statistiques permettent de répondre à un certain type de questions, mais que seul un expert du domaine d'application est capable de formuler une question pertinente et d'en interpréter les résultats. Ainsi, dans cette thèse les méthodes sont générales mais leur utilisation est faite dans le cadre du problème industriel initialement posé : l'estimation de l'erreur de modèle présente dans les simulations diphasiques en Mécanique des Fluides Numérique (CFD) dans le contexte nucléaire.

Contexte industriel : simulations CFD dans le domaine nucléaire

Codes système en thermo-hydraulique

La simulation numérique est un outil essentiel à l'étude de la variété de scénarios qui peuvent se produire lors du fonctionnement de centrales nucléaires. Nous nous intéressons particulièrement à la thermo-hydraulique, qui concerne l'étude des écoulements de fluides comprenant des transferts de chaleur importants, pouvant impliquer des phénomènes thermodynamiques comme du changement de phase. La plupart des scénarios concernés sont dits "multi-échelles", car ils requièrent la bonne représentation de phénomènes se produisant à différentes échelles spatiales. Prenons l'exemple d'un écoulement se produisant dans un générateur de vapeur, où une forte ébullition se produit : à la plus fine échelle, il serait nécessaire de représenter chaque cavité micrométrique de la surface d'ébullition; à l'échelle du millimètre, chaque bulle individuelle pourrait être simulée; et à l'échelle du mètre, c'est l'écoulement diphasique entier qu'il est nécessaire de calculer pour étudier son débit, ou l'usure thermique induite sur les matériaux.

Il est inenvisageable de représenter directement l'ensemble de ces phénomènes dans une unique simulation numérique, car le raffinement spatial requis nécessiterait un trop grand nombre de cellules de calcul. Par conséquent, il existe des codes dédiés à chaque usage, en fonction de la taille du domaine de simulation. Les codes système permettent de simuler l'ensemble des composants d'un circuit (réacteur, turbines, pompes) et leurs interactions. Des exemples de tels codes sont RELAPS5, TRACE[1], ou CATHARE[2] pour le CEA. Les codes à l'échelle composant permettent de représenter fidèlement les écoulements à l'intérieur de composants spécifiques comme le coeur du réacteur (code FLICA5) ou les générateurs de vapeur (code GENEPI+). L'échelle la plus fine, permettant aussi bien de simuler un mélange de fluides froid et chaud que de l'ébullition en paroi, est appelée l'échelle locale. Les codes utilisés pour simuler des écoulements à cette échelles sont appelés les codes de CFD. De nombreux solveurs commerciaux ou de recherche sont disponible à cet effet, par exemple TRIO_CFD[3], NEPTUNE_CFD[4] ou Ansys Fluent[5].

[1] https://www.nrc.gov/about-nrc/regulatory/research/safetycodes.html
[2] https://cathare.cea.fr/
[3] https://triocfd.cea.fr/
[4] https://www.code-saturne.org/cms/web/NEPTUNECFD
[5] https://www.ansys.com/fr-fr/products/fluids/ansys-fluent

Prenant l'exemple du code CATHARE, l'écoulement des fluides est simulé à partir de six équations [Bestion, 1990] qui correspondent aux équations de conservations de la masse, quantité de mouvement et énergie pour les phases liquide et vapeur. Chacune de ces équations fait intervenir des lois de fermeture, qui représentent les échanges de ces trois quantités entre les phases mais également avec les parois. Ces lois sont établies à partir de modélisations physiques car elles doivent être en mesure de représenter un éventail de phénomènes complexes seulement à partir de grandeurs moyennées, étant donné la large taille de maille spatiale dans ces calculs. Par conséquent, elles contiennent un fort risque de présence de erreur de modèle qui doit être contrôlée par le biais d'études de Vérification, Validation et Quantification d'Incertitudes (V&V&UQ), dont nous préciserons la nature dans la section suivante.

Une précision importante est que notre travail ne porte pas directement sur les lois de fermeture utilisés dans les codes système, mais plutôt sur celles utilisées dans les codes CFD, dont les résultats servent de référence pour évaluer les calculs à l'échelle composant ou système. Nous ferons la distinction entre trois échelles de CFD. La plus fine échelle est appelée Simulation Numérique Directe (DNS), où les équations de Navier-Stokes sont résolues directement sur le maillage. Elle est par exemple adaptée à la simulation d'essaims de bulles, requérant une résolution d'environ 10 mailles par diamètre de bulles. Ensuite vient la Simulation des Grandes Structures de la turbulence (LES), qui font intervenir des mailles plus grandes où les effets de turbulence sont modélisés. A la plus grande échelle sont les approches Navier-Stokes par Moyenne de Reynolds (RANS), où les équations résolues sont des équations moyennées, ce qui implique d'avoir recours à un plus grand nombres de lois de fermeture. L'approche moyennée est une solution classique pour les écoulements diphasiques pour éviter de simuler directement les interfaces, leurs mouvements et déformations (bulles, poches, gouttes, etc.). Elle fonctionne également sur le principe de bilans de conservation associés à des lois de fermeture, démontrant la nécessité de considérer l'erreur de modèle à leur égard.

Vérification, Validation et Quantification d'Incertitudes

Avant de pouvoir accepter les résultats obtenus à partir d'une simulation numérique comme des éléments scientifiques solides sur lesquels s'appuyer dans les études de conception ou de sûreté, il est nécessaire de fournir des preuves de leur pertinence par des étapes de V&V. La première étape concerne la détermination des erreurs d'origine numérique (raffinement de maillage, ou bugs dans le code) et leur réduction. La seconde étape consiste à comparer les résultats de simulations à des données expérimentales représentatives du phénomène physique étudié, pour confirmer leur validité.

Les étapes de V&V ont été employées dès le début de l'utilisation des codes système dans les années 1950, et elles sont d'autant plus importantes aujourd'hui. Leurs spécifications n'ont cessé d'évoluer pour s'adapter à la complexité croissante des systèmes et des outils numériques employés. Elles sont généralement précisées sous la forme de normes, spécialement édictées selon le domaine d'application. Concernant la CFD, des normes appliquées aujourd'hui sont par exemple [AIAA, 1998] et [ASME, 2009], provenant respectivement de l'Institut Américain d'Aéronautique et d'Astronautique, et de la Société

Américaine des Ingénieurs en Mécanique. Une troisième étape, plus récente, apporte un complément aux étapes de V&V : il s'agit de la Quantification d'Incertitudes (UQ). Son principe est d'estimer les diverses sources d'incertitudes dans une simulation numérique et d'évaluer leur impact sur les prédictions. Les études UQ s'inscrivent naturellement dans une démarche émergente de simulation : la méthode BEPU [D'Auria et al., 2012]. L'idée de cette démarche est de fournir des résultats de calcul probabilistes avec des intervalles de crédibilité associés (équivalent bayésien des intervalles de confiance), en opposition avec l'approche traditionnelle conservative qui vise à garantir la sûreté par des facteurs de sécurité importants, pouvant conduire à un surdimensionnement des systèmes. La méthodologie BEPU a été explorée lors de projets internationaux comme PREMIUM [CSNI, 2016], puis SAPIUM [Baccou et al., 2020].

En thermo-hydraulique, une manière classique de procéder est la Quantification Directe d'Incertitudes (fUQ), dans laquelle une incertitude sur les paramètres est formulée, puis son impact sur la prédiction est évalué en réalisant des simulations. C'est par exemple l'approche de la méthode CSAU [Boyack et al., 1990; Zhao and Mousseau, 2012] ou GRS [Glaeser, 2008]. Une autre manière de procéder est par la Quantification Inverse d'Incertitudes (iUQ), dans laquelle l'incertitude sur les paramètres est progressivement réduite par l'acquisition de données expérimentales et leur comparaison aux résultats de simulation, appartenant à la famille des problèmes inverses. Une revue des méthodes iUQ appliquées aux codes système est proposée dans [Wu, Xie, et al., 2021], dans laquelle l'auteur distingue trois catégories de méthodes. Les méthodes fréquentistes ont pour objectif de déterminer des valeurs optimales de paramètres (ex: CIRCE [De Crecy and Bazin, 2004]). Les méthodes bayésiennes représentent les paramètres comme des variables aléatoires (ex: Approche Bayésienne modulaire [Wu, Shirvan, et al., 2019]), et les méthodes empiriques déterminent parmi un ensemble de valeurs de paramètres celles qui répondent à des critères préétablis (ex: DIPE [Joucla and Probst, 2008]).

Les méthodes iUQ requièrent donc l'utilisation de données expérimentales, aussi le choix des expériences à réaliser, leur nombre et leur conception sont centraux pour une détermination précise des incertitudes. En raison du caractère multi-échelle et multi-physique des phénomènes rencontrés en thermo-hydraulique, l'approche privilégiée est l'utilisation de Tests à Effets Séparés (SET), pour établir de nouvelles lois de fermeture pertinentes et évaluer les existantes. Ces tests sont conçus pour être chacun représentatif d'un phénomène physique spécifique, sans interactions avec d'autres. En général, les expériences associées sont à échelle réduite en raison de la taille importante du système réel, ce qui est source d'un nouveau type d'incertitude: l'effet d'échelle. Cet effet, non traité dans le présent manuscrit, est décrit par exemple dans [D'Auria et al., 2012; Ivanov et al., 2018]. Les Tests à Effets Intégraux (IET) sont eux représentatifs d'une partie, ou de l'ensemble, des phénomènes physiques représentés dans l'application. En raison du coût requis pour proposer des expériences adaptées, les expériences sur lesquelles s'appuient la Validation et UQ restent en nombre trop limité pour les ingénieurs. Il est également difficile de mettre en place une instrumentation sur les réacteurs utilisés en production, et pour des raisons évidentes on ne génère pas volontairement des accidents pour en étudier les conséquences.

Les simulations CFD, dans ce contexte, ont le potentiel de servir d'"expériences numériques" pour fournir des données de références qui viendraient nourrir les études V&V&UQ. Elles possèdent une importance moindre par rapport aux données expérimentales réelles, mais la maturité croissante des outils numérique renforce petit à petit leur place dans les études de conception et de sûreté, principalement en raison de leur coût bien moins important. Il est donc nécessaire d'évaluer correctement l'erreur contenue dans les résultats de simulations CFD, qui sont des éléments essentiels à une conception plus sûre des réacteurs d'aujourd'hui et de demain.

Enjeux et contributions

Enjeux

Proposer une nouvelle méthode UQ dans ce contexte fait intervenir à la fois des contraintes liées à l'objectif de la méthode, et son cadre d'application particulier. La méthode proposée doit permettre de :

- **Quantifier les incertitudes de modèle** qui s'appliquent aux lois de fermeture employées (ex : interactions entre les phases, tranferts de chaleur, ébullition), et évaluer leur impact sur les prédictions. Leur détermination est particulièrement difficile lorsque la prédiction initiale est une représentation peu satisfaisante du phénomène physique, et qu'elle est fortement sensible aux valeurs des paramètres utilisés.
- **Quantifier les incertitudes expérimentales** qui jouent un rôle décisif pour les méthodes iUQ. Il est souvent difficile de formuler des hypothèses adaptées à leur représentation, et les résultats des méthodes iUQ y sont souvent sensibles.
- **Trouver des valeurs appropriées pour les paramètres des modèles**, en prenant en compte les deux sources d'incertitude mentionnées. Il est nécessaire de déterminer de bons paramètres et de connaître la plage de fonctionnement dans laquelle le modèle est valide. Si les deux sources d'incertitudes précédentes ne sont pas correctement estimées, les valeurs de paramètres obtenues peuvent être mauvaises et fausser les conclusions de l'étude.
- **Mener des études sur une large base de données expérimentale**. Il est nécessaire de disposer d'un panel d'expériences pour que la Validation et UQ soient pertinentes. La plage d'utilisation d'un modèle doit être en théorie incluse dans celle engendrée par les données dont on dispose. Utiliser une base trop réduite fait courir le risque d'*overfit* de modèle, et d'obtenir des résultats non valides en général.
- **Identifier les amélioration possibles d'un modèle physique**. Dans le cas où, après application des méthodes UQ, l'erreur de modèle serait déterminée trop importante pour l'utilisation envisagée, il est désirable de pouvoir donner des éléments qui souligneraient les parties du modèle à améliorer.

De plus, travailler sur des codes CFD dans le nucléaire pose les contraintes suivantes

:

- **Les codes présentent un très grand nombre de paramètres**. Les codes CFD multiphasiques emploient des dizaines de modèles ou corrélations expérimentales pour représenter des phénomènes physiques, conduisant parfois à plusieurs centaines de coefficients et paramètres. Les méthodes UQ ne peuvent pas considérer la totalité de ces paramètres en raison du *fléau de la dimension*, qui fait croître de manière exponentielle le coût de calcul en fonction de la dimension. Il est donc nécessaire de trier parmi ces paramètres, soit par des arguments physiques ou statistiques.
- **Les codes sont coûteux à évaluer**. Ces simulations sont généralement réalisées sur serveur de calcul, en bénéficiant d'une parallélisation forte. Dans notre seconde application, une simulation coûte deux heures de calcul sur 30 processeurs. Il est donc nécessaire de limiter le nombre de simulations réalisées pour des raisons de budget. Cependant, les méthodes UQ requièrent le plus de simulations possibles pour obtenir une bonne précision sur les résultats. Il est donc nécessaire d'optimiser les simulations à réaliser, à budget fixé.
- **Les codes ont un comportement fortement non-linéaire**. En raison de la complexité des solveurs, les résultats de simulation dépendront de manière non-linéaire des paramètres ou coefficients. Ainsi, de nombreuses techniques UQ qui fonctionnent sur le principe d'un développement au premier ordre de cette quantité ne seront pas utilisables dans ce contexte. Les non-linéarités ne peuvent être appréciées que par un nombre suffisant de simulations du code, rejoignant la contrainte du point précédent.
- **Des erreurs d'origine numérique peuvent être présentes**. Les simulations sont généralement non-stationnaires, ce qui demande d'estimer avec précision les erreurs liées à la convergence en temps du calcul, en plus de la discrétisation spatiale du maillage. Ces deux sources d'erreur pouvant être confondues avec de l'erreur de modèle, il est nécessaire de les contrôler par des études supplémentaires.

En réponse aux enjeux mentionnés, les contributions apportées dans ce travail sont détaillées dans les paragraphes ci-dessous :

Nouvelle méthode de calibration avec erreur de modèle

Définition plus générale de l'erreur de modèle

Notre point de départ a été la remise en question de la définition usuelle de l'erreur de modèle, proposée par Kennedy et O'Hagan [Kennedy and O'Hagan, 2001a], dans laquelle elle est établie comme la différence entre le phénomène décrit et les prédictions du modèle obtenues à la "véritable valeur" de ses paramètres d'entrée. Nous avons identifié cette définition comme la source du problème d'identifiabilité du terme d'erreur de modèle dans la formulation originelle du problème. Ce problème a été longuement caractérisé dans la littérature [Gustafson, 2005; Brynjarsdóttir and O'Hagan, 2014; Maupin and Swiler, 2020] comme étant un point d'attention de la méthode KOH. L'idée est que, d'après la

définition de l'erreur de modèle, seule la "véritable valeur" des paramètres du modèle peut nous permettre de la déterminer. Cependant, puisqu'elle est inconnue, plusieurs formes d'erreur de modèle peuvent survenir, et il peut être difficile de distinguer la plus probable. Notons également que, mathématiquement, l'estimation KOH de l'erreur de modèle fait intervenir l'ensemble des valeurs des paramètres par leur densité a priori (eq. (2.21)), ce qui va à l'encontre de leur définition. Plusieurs approches cherchent à résoudre ce problème : par exemple, Arendt [Arendt, Apley, Chen, et al., 2012; Arendt, Apley, and Chen, 2012] privilégie l'utilisation d'un grand nombre de données expérimentales, et d'un modèle à valeurs vectorielles, pour discerner la bonne quantité. Des méthodes de correction du modèle sont proposées par [Ling et al., 2014] pour déterminer les paramètres identifiables, à partir de la Matrice d'Information de Fisher. Cette approche a été étendue par [Kim and Youn, 2019] aux modèles hiérarchiques. L'"approche modulaire" de [Liu et al., 2009] permet de diviser le problème joint d'estimation en des modules séparés, pour améliorer l'identifiabilité de chaque module. Nous avons examiné quelques méthodes de la littérature sous cet angle. Abordant le problème de manière fréquentiste, Tuo [Tuo and Wu, 2015a] enrichit le processus de calibration en définissant la meilleure valeur de paramètres comme celle qui minimise la norme L2 au phénomène décrit. Plumlee [Plumlee, 2017] propose une approche bayésienne partant de la même définition.

La définition externe de la véritable valeur des paramètres, si elle permet de résoudre le problème d'identifiabilité, ne nous a pas paru satisfaisante car nous travaillons justement dans le cas où les valeurs "best fit" peuvent être trompeuses, en raison de la présence d'une erreur de modèle forte. Le problème d'identifiabilité prend son origine du fait que, pour chaque valeur de paramètres du modèle, une distance aux observations peut être calculée, alors qu'une seule est valide d'après la définition KOH. De plus, cette "véritable valeur" dépend des hypothèses de calibration (ex: fonction de covariance de l'erreur de modèle), ce qui est reconnu par Kennedy et O'Hagan dans leur article, mais qui est vu par d'autres auteurs comme un problème à corriger [Tuo and Wu, 2015a]. Notre idée est donc de définir l'erreur de modèle comme la différence entre le phénomène décrit et la moyenne a posteriori des prédictions du modèle, après calibration. Cette définition permet de faire apparaître de manière explicite la dépendance aux hypothèses de calibration, mais surtout à l'ensemble des valeurs des paramètres du modèle.

Nouveau cadre de calibration et approximation FMP

Notre nouvelle définition permet d'établir une équation de calibration avec dépendance explicite de l'erreur de modèle à la valeur des paramètres du modèle. Cette équation a déjà été étudiée par Plumlee [Plumlee, 2017], et est similaire à celle de la *Full Bayesian Analysis* [Higdon, Kennedy, et al., 2004; Higdon, Gattiker, et al., 2008], où est effectuée l'estimation jointe de paramètres et hyperparamètres. L'équation peut être résolue directement par des méthodes bayésiennes classiques (échantillonnage de la densité a posteriori par MCMC), ce qui constituera dans tout le manuscrit la solution de référence au problème de calibration, dénommée "solution Bayes". Cependant, elle nécessite d'échantillonner sur l'espace joint paramètres-hyperparamètres, qui peut être de grande dimension selon l'application considérée, engendrant des problèmes de qualité de

mélange des chaînes et de temps de calcul trop important.

Nous avons proposé l'approximation FMP, en introduisant une relation fonctionnelle entre chacune de ces deux quantités. Les hyperparamètres sont estimés non plus par le maximum de leur densité a posteriori (KOH), mais par le maximum de la densité conditionnelle à chaque valeur de paramètres. L'approche KOH est de comparer l'ensemble des prédictions en utilisant une unique forme d'erreur de modèle déterminée à l'avance. Notre approche consiste à établir pour chaque prédiction la forme d'erreur de modèle qui lui est la plus favorable, et de comparer les prédictions entre elles par le score (valeur de vraisemblance) correspondant à leur forme d'erreur de modèle. Les hypothèses nécessaires pour que la solution FMP soit une bonne approximation de la solution de référence sont données, et reposent sur le bon choix de fonction de covariance du terme d'erreur de modèle. Des bonnes propriétés de la méthode, à savoir la continuité de la fonction d'approximation et l'hémi-continuité des hyperparamètres optimaux, sont données. On démontre que le critère à optimiser est dérivable et ses dérivées sont explicitées, permettant d'utiliser des méthodes à gradients pour résoudre les optimisation plus rapidement.

Notre approche est plus coûteuse que l'approche KOH, car plus générale : elle nécessite un plus grand nombre d'estimations d'hyperparamètres, obtenues par résolution d'un problème d'optimisation. Il est alors nécessaire de démontrer l'avantage de l'utiliser par rapport à la méthode KOH. Pour ce faire, nous prenons une hypothèse de normalité sur la densité jointe a posteriori du problème de référence. Premièrement en tant que gaussienne unimodale, puis en tant que mélange de gaussiennes à modes à projections séparées, nous démontrons deux points de défaillance de la méthode KOH. Le premier point est qu'elle sous-estime systématiquement la variance de la densité a posteriori des paramètres, induisant un effet de "fausse confiance", et des intervalles de crédibilité sous-estimés autour des prédictions finales. Ainsi, la méthode KOH sous-estime l'incertitude paramétrique et ses effets sur les prédictions. Deuxièmement, dans le cas multimodal, l'estimation KOH peut ignorer un mode qui correspondrait à une interprétation alternative des données. Cela a pour effet de manquer une zone de l'espace contenant de la probabilité a posteriori, et aboutir à des conclusions faussées. En comparaison, la méthode FMP estime fidèlement les variances et retrouve l'ensemble des modes de la densité de référence. De par l'approximation utilisée, elle peut inverser l'importance relative des modes, ce qui peut influencer les prédictions moyennes mais a généralement peu d'effet sur les intervalles de confiance finaux. Nous proposerons par la suite un algorithme de rééchantillonnage par importance pour corriger cette estimation biaisée des pics de densité.

Etudes numériques d'estimation dans un régime asymptotique

Nous avons construit deux études numériques pour traiter deux questions sous-jacentes à notre étude précédente: premièrement, est-il pertinent de faire une approximation de Dirac des densités conditionnelles ? i.e. dans les exemples courants d'application, la densité conditionnelle est-elle plate ou piquée ? et deuxièmement, l'approximation de normalité de la densité jointe a posteriori est-elle pertinente ?

Dans la première étude nous montrons que, dans la plupart des fonctions de covariance utilisés pour la calibration, les hyperparamètres sont non-microergodiques, une propriété mathématique qui établit que plusieurs valeurs d'hyperparamètres peuvent engendrer des mesures gaussiennes équivalentes [Stein, 1999; Bachoc, 2014]. Cette propriété se traduit dans un exemple concret par un plateau de la fonction de vraisemblance autour de la véritable valeur d'hyperparamètres, correspondant aux autres valeurs équivalentes d'hyperparamètres. L'approximation de Dirac n'est donc pas adaptée dans cette situation. Cependant, cela a peu d'influence sur le résultat principal de la calibration, qui est la densité a posteriori des paramètres : en effet, le score FMP d'une valeur de paramètres est fonction de la valeur de la fonction de vraisemblance, ainsi puisqu'elle est relativement plate autour du véritable maximum, la calibration est tout de même correctement effectuée.

Pour la seconde étude nous commençons par introduire trois cadres asymptotiques d'acquisition d'observations expérimentales. L'asymptotique à domaine fixe consiste à densifier le nombre d'observations dans un intervalle restreint, alors que l'asymptotique à domaine croissant consiste à acquérir progressivement des observations hors du domaine initial. Ces deux asymptotiques sont régulièrement étudiées dans la littérature scientifique portant sur le krigeage. La troisième, que nous nommons asymptotique à domaine répété, consiste à acquérir de nouvelles observations toujours aux mêmes points, mais qui seront a priori différentes si l'erreur de mesure est importante. Nous étudions deux exemples élémentaires, l'un où le modèle est correctement spécifié (il existe une valeur de paramètres qui permet de reproduire exactement la réalité), et un cas incorrectement spécifié. Les trois régimes asymptotiques sont appliquées dans ces deux cas. Deux résultats principaux sont à retenir de cette étude: premièrement, les résultats varient peu selon les trois régimes, et les bons paramètres quand ils existent sont correctement identifiés (propriété de consistance). Deuxièmement, la densité des hyperparamètres ne peut pas être approchée par une gaussienne en raison de supports tronqués et de possibles problèmes de non-identifiabilité entre variance et longueur de corrélation.

Techniques numériques d'approximation

Chaînes de Markov Monte-Carlo appliquées en calibration

La méthode FMP présentée dans ce manuscrit est plus coûteuse que la méthode KOH car elle offre une plus grande précision. Il est donc d'intérêt de proposer des algorithmes efficaces pour réaliser la calibration avec erreur de modèle, et d'employer des solutions pour réduire le coût de la méthode FMP.

Nous utilisons une technique standard en inférence bayésienne, l'échantillonnage de la densité a posteriori par MCMC, plus précisément en suivant l'algorithme de Metropolis-Hastings. Nous détaillons la manière dont cet algorithme se décline suivant les trois méthodes de calibration étudiées, KOH, FMP et Bayes. Nous donnons également quelques techniques de diagnostic de convergence de chaînes, appliquées à une chaîne ou plusieurs, qui seront utilisées dans les chapitres d'application. Enfin, nous proposons un algorithme de rééchantillonnage dont le principe est d'améliorer la qualité de l'échantillon FMP pour

qu'il soit une meilleure approximation de l'échantillon Bayes, au prix de réduire sa taille par un facteur 20. Cet algorithme est illustré sur un exemple numérique élémentaire avec modèle mal spécifié.

Construction de modèles de substitution

Les techniques d'accélération proposées dans ce manuscrit sont basées sur la construction de modèles de substitution, ou méta-modèles. Un premier méta-modèle est utilisé pour le code de calcul à calibrer, dont les évaluations sont généralement coûteuses. Dans la situation où la sortie du code est de grande dimension, comme par exemple si la prédiction est une courbe au lieu d'une quantité ponctuelle, il est judicieux de se tourner vers des techniques de réduction de dimension. C'est l'objectif de l'Analyse par Composantes Principales (ACP), qui est utilisée pour déterminer les modes propres de la sortie du code, ainsi un modèle de substitution peut être construit par combinaison linéaire de ces modes, les coefficients étant remplacés par des Processus Gaussiens. La sélection du nombre de modes choisis s'opère par évaluation de la performance du méta-modèle sur un ensemble d'observations non utilisé pour l'apprentissage, qui constitue l'ensemble de validation.

Il est également possible d'employer des méta-modèles dans l'application de la méthode FMP. Rappelons que le coût principal provient du fait que, dans l'algorithme de Metropolis-Hastings, chaque étape requiert un calcul des hyperparamètres optimaux (optimisation) puis une évaluation de la fonction de vraisemblance. Le nombre d'étapes pouvant être important (de l'ordre de 10^4 à 10^6 pour nos applications), l'utilisation de méta-modèles est presque indispensable pour des applications concrètes. Une première possibilité est de chercher à apprendre la fonction de vraisemblance FMP : à partir de quelques valeurs de paramètres de modèle choisies, il est possible de construire une fonction approchée qui sera utilisable dans l'algorithme de Metropolis-Hastings, chaque étape de cet algorithme nécessitera ainsi seulement une évaluation du méta-modèle. Une deuxième possibilité est de construire plusieurs méta-modèles sur les hyperparamètres optimaux, permettant de remplacer l'étape d'optimisation. Nous démontrons sur un exemple simple que la précision obtenue par ces deux méthodes est comparable. L'avantage de la première méthode est que la fonction-cible est continue, et qu'il est préférable de construire un unique méta-modèle plutôt qu'un nombre égal aux nombre d'hyperparamètres du problème. La seconde méthode présente l'avantage d'offrir l'accès aux distributions a posteriori des hyperparamètres, ainsi que de pouvoir effectuer des prédictions corrigées du modèle sans optimisation supplémentaires. Ainsi, la seconde méthode sera préférée dans les chapitres d'application. Notons que, dans le chapitre 5, les méta-modèles seront utilisés pour fournir le point de départ d'une optimisation des hyperparamètres dans l'algorithme Metropolis-Hastings. Par conséquent, le coût de la méthode FMP est réduit mais la précision reste exacte.

Choix itératif des points d'apprentissage

Nous nous intéressons ensuite à la construction des méta-modèles des hyperparamètres optimaux, plus précisément aux choix des points d'apprentissage où seront réalisées des

optimisations. Une approche naïve est de placer ces points selon un plan d'expériences numérique à remplissage d'espace, type Latin Hypercube Sampling ou Quasi Monte-Carlo. Les points sont fortement espacés dans le but de représenter correctement l'espace a priori des paramètres. Une stratégie alternative est de placer ces points dans la zone de probabilité a posteriori des paramètres, ce qui permet de raffiner la précision des méta-modèles dans les zones qui seront fortement explorées lors de l'étape de MCMC. Nous démontrons que la précision des méta-modèles contrôle le degré d'approximation de la postérieure vers la postérieure FMP. Ainsi, cette stratégie est justifiée, mais les zones probables a posteriori ne sont pas connues à l'étape initiale de création des méta-modèles. Il est donc nécessaire de recourir à une procédure itérative.

Notre approche est de commencer par une approximation grossière de la fonction de vraisemblance FMP, en commençant par quelques évaluations tirées selon un plan à remplissage d'espace. Ensuite, un algorithme MCMC est appliqué pour obtenir un échantillon de cette approximation grossière. De cet échantillon on obtient un ensemble de nouveaux points d'apprentissage, par une technique de rééchantillonnage dont les poids sont proportionnels à la variance de prédiction des méta-modèles, qui est représentative de l'erreur de prédiction en ces points. Ainsi, notre algorithme suit le principe d'équilibre entre exploration (choisir des points éloignés du training set actuel, où la variance de prédiction est grande) et exploitation (choisir des points où la densité a posteriori est supposément grande, car ils sont sélectionnés dans un échantillon de cette loi). Nous démontrons que cette approche permet, à nombre de points d'apprentissage égaux, d'obtenir une précision meilleure que les plan d'expériences à remplissage d'espace, la précision étant calculée sur un échantillon de la véritable postérieure FMP. Cet algorithme de construction itérative des méta-modèles des hyperparamètres optimaux sera un point clé dans les chapitres d'application.

Calibration d'un modèle de répartition de flux de chaleur

Contexte

La première application concerne un modèle de calcul de flux de chaleur, entre une paroi et un liquide, dans des conditions d'ébullition allant de la première apparition de bulles à un régime fortement turbulent. Le modèle étudié est le MITB [Kommajosyula, 2020], qui prend comme données d'entrée la vitesse du fluide, sa température, la pression, la géométrie du système et la température de la paroi. Le modèle est semi-empirique, mettant en lien des corrélations établies expérimentalement, en fonction de considérations physiques. Nous l'avons directement implémenté dans le langage C++, et son évaluation est quasiment instantanée. Le modèle fait intervenir entre autres une corrélation expérimentale portant sur le diamètre de bulle à leur départ de la paroi. Les coefficients de la corrélation ont été établis par Kommajosyula à partir de mesures de diamètres de bulles dans différentes conditions d'écoulement. Notre approche est de calibrer deux coefficients de cette corrélation en utilisant des données de flux de chaleur obtenues dans les expériences de Kennel [Kennel, 1949]. Ainsi, il sera possible de donner des barres d'erreur sur les prédictions de flux et de diamètre, permettant une représentation plus fidèle des données. De plus, nous intégrons dans la calibration la valeur de l'angle de

contact entre le liquide et la paroi, qui est fortement dépendante de l'état de surface de la paroi et des conditions de pression et température. Cette valeur n'ayant pas été mesurée dans l'expérience de Kennel, Kommajosyula a choisi $\theta_c = 40°$ pour les reproduire, mais en vérité il existe une incertitude de plusieurs dizaines de degrés sur cette valeur. La calibration avec prise en compte de l'erreur de modèle a donc pour but d'apprendre fidèlement ces trois paramètres, et d'améliorer les prédictions du MITB sur les expériences de Kennel ainsi que du diamètre de bulle.

La difficulté principale de cette application réside dans le fait que les expériences Kennel sont au nombre de 13 et qu'il est nécessaire d'estimer un terme d'erreur de modèle pour chacune, portant le nombre d'hyperparamètres à 26, ce qui est considérable pour une telle application. De plus, l'erreur de mesure porte principalement sur une entrée du modèle (la température de la paroi) et non sur sa sortie, ainsi elle sera estimée dans une étape préliminaire et ne sera pas apprise au même titre que les autres hyperparamètres. Nous menons une première étude de calibration en considérant chaque expérience de manière individuelle, obtenant ainsi pour chaque expérience la meilleure représentation possible des données. Cela permet de séparer les expériences en trois groupes où l'erreur de modèle est soit faible, soit moyenne, soit importante. Nous menons ensuite deux calibrations, la première avec les 9 expériences où l'erreur de modèle n'est pas importante, et la seconde avec l'ensemble des expériences. Ces calibrations aboutissent à une bonne représentation globale des données. La procédure permet d'apprendre correctement les trois paramètres étudiés et génère des intervalles de crédibilité sur les prédictions du modèle qui permettent de recouvrir la plupart des points expérimentaux, en flux et en diamètre.

Résultats de calibration

Les comportements des méthodes KOH, FMP et Bayes sont étudiés. Sur les calibrations individuelles, lorsque l'erreur de modèle est faible, les trois méthodes coïncident. Pour les autres groupes d'expériences, les intervalles de crédibilité estimés par la méthode KOH sont systématiquement plus restreints que ceux obtenus par FMP et Bayes. Ceux obtenus par FMP sont plus proches des résultats Bayes. Cela confirme que l'effet de fausse confiance est atténué par notre méthode. Les postérieures des paramètres obtenues par KOH présentent des variances plus faibles et celles des hyperparamètres sont des Diracs, alors que les distributions obtenues par FMP représentent plus fidèlement la variance a posteriori dans les deux cas. Les conclusions obtenues sur les paramètres sont différentes pour chaque expérience, car à chacune correspond des valeurs optimales. Il est donc nécessaire de considérer l'ensemble des expériences pour donner des conclusions pertinentes sur les paramètres.

La calibration avec les expériences des groupes avec faible et moyenne erreur de modèle est un cadre où la solution Bayes est atteignable par MCMC, avec un mélange long à obtenir en raison de la grande dimension du problème. La méthode KOH, appliquée en utilisant les hyperparamètres calculés sur les expériences individuelles, donne des résultats incohérents. Il est nécessaire de l'adapter en calculant des erreurs de modèles qui prennent en compte l'ensemble des expériences, méthode que nous nommons "KOH

pooled". Nous montrons donc qu'il est impossible de suivre, selon la méthode KOH, une démarche itérative où plusieurs expériences seraient incluses progressivement, car il est en fait nécessaire de recalculer l'erreur de modèle de chacune lorsque de nouvelles données sont intégrées. De plus, l'optimisation à réaliser croît en dimension à chaque rajout, ce qui augmente la complexité du problème. En comparaison, les hyperparamètres optimaux de la méthode FMP obtenus sur chaque expérience sont directement réutilisables pour une calibration globale. Ainsi, les dimensions des optimisations FMP restent faibles (dimension 2) et ne dépendent pas du nombre d'expériences considérées. Les résultats de la méthode KOH pooled, bien que les hyperparamètres soient correctement estimés au maximum de leur véritable postérieure, sont moins proches de la solution Bayes que ne le sont les résultats FMP, et montrent toujours l'effet de fausse confiance.

En considérant l'ensemble des 13 expériences Kennel, la solution Bayes n'est plus atteignable dans un temps de calcul raisonnable. Seuls les résultats FMP et KOH pooled peuvent être comparés. Il est remarquable de constater que les résultats FMP diffèrent peu de la calibration avec 9 expériences, et donc des résultats de référence obtenus précédemment. Les résultats KOH pooled sont eux centrés autour d'une autre valeur de paramètres, avec une variance suffisamment grande pour inclure autant de points expérimentaux que FMP dans les intervalles de crédibilité autour des prédictions. Les deux méthodes fournissent donc deux explications différentes des données, sans solution de référence pour les départager. Le taux de couverture, en tant que proportion de l'intervalle de confiance associé aux mesures qui est inclus dans les intervalles de crédibilité autour des prédictions est favorable à KOH pooled, mais la méthode FMP fournit des prédictions corrigées qui sont plus réalistes d'un point de vue physique que KOH pooled. La proximité entre les résultats FMP et les résultats Bayes précédents montre que l'inclusion d'expériences avec grande erreur de modèle n'influence que peu la méthode FMP, ce qui est un résultat de robustesse désirable.

Le chapitre 5 a donc été l'occasion de démontrer l'avantage d'utiliser la méthode FMP lorsque l'on souhaite calibrer un modèle en utilisant un grand nombre d'expériences et un terme d'erreur de modèle pour chacune. Les résultats sur les paramètres indiquent que la corrélation expérimentale pour le diamètre de bulle, telle qu'initialement calibrée, avait sous-estimé l'effet de la température de la paroi. De plus, l'angle de contact le plus consistent avec les observations est de l'ordre de 10°, ce qui peut être expliqué par le fait que la surface de l'élément chauffant dans l'expérience de Kennel était en réalité plus rugueuse qu'estimé.

Calibration du solveur NEPTUNE_CFD pour la simulation d'un écoulement diphasique

Contexte

La seconde application présentée dans le chapitre 6 concerne un solveur CFD multiphasique, NEPTUNE_CFD, et les expériences DEBORA qui reproduisent l'écoulement de bas en haut d'un fluide dans une conduite verticale, avec parois chauffantes causant une ébullition qui est progressivement plus importante selon la hauteur dans la conduite. Nous sélectionnons deux quantités d'intérêt dans ces expériences : le taux de vide (pro-

portion volumique de gaz dans le mélange) et le diamètre des bulles. Le profil radial de ces deux quantités est mesuré dans une section de la conduite, située à une hauteur importante, là où l'ébullition est la plus forte. Les expériences sont réalisées dans différentes conditions de pression, flux chauffant, débit et température liquide. Les données expérimentales sont prises d'une seule expérience, dénommée cas A6. En raison du budget de simulation alloué, il a été privilégié d'utiliser une seule expérience et de pouvoir explorer finement l'espace des paramètres calibrés, plutôt que d'accroître la taille de la base expérimentale et perdre en précision sur les paramètres.

Le constat de départ est que le code NEPTUNE_CFD prédit mal les deux quantités d'intérêt considérées sur le cas A6 : le taux de vide est sur-prédit (40% en paroi pour une valeur mesurée de 30%), le diamètre de bulle est correctement prédit en paroi mais sous-prédit d'un facteur 5 au centre de la conduite. Certains modèles physiques utilisés dans le calcul sont susceptibles d'être la raison de ces mauvaises prédictions. L'objectif de l'étude présentée dans ce chapitre est donc d'évaluer l'erreur présente dans les modèles en question, et d'introduire des modifications simples des modèles pour chercher à améliorer les prédictions des quantités d'intérêt.

Etude préliminaire

La première partie de l'étude est consacrée à la formulation de différentes hypothèses physiques pouvant expliquer l'erreur dans les prédictions. Trois pistes sont explorées : des incertitudes expérimentales importantes, un modèle de répartition de flux de chaleur qui surestimerait l'évaporation, ou le champ d'aire interfaciale qui serait incorrectement transporté. Nous étudions ces trois pistes par des analyses OAT, où la sensibilité du code à ces hypothèses est estimée de manière simple en réalisant quelques calculs. Le modèle de transport de l'aire interfaciale est modifié en introduisant des coefficients multiplicateurs devant les termes source qui représentent des phénomènes physiques concernant les bulles dans l'écoulement (fragmentation, coalescence, nucléation, ...). Nous avons choisi ce type de modification en raison de sa simplicité de mise en place, et du point de vue d'estimer l'erreur dans la modélisation existante de ces termes sans chercher à proposer de nouveaux modèles. A la suite des analyses OAT, le cadre de calibration est établi : quatre coefficients multiplicateurs seront introduits devant les termes source les plus importants, en sus d'une incertitude expérimentale sur le flux chauffant à la paroi. Les domaines de variation des coefficients sont suffisamment restreints pour ne pas dénaturer le modèle original. La variation de l'incertitude expérimentale est également établi en fonction de l'incertitude de mesure reportée dans l'expérience.

Résultats de calibration

Une étape préliminaire à la calibration est la construction d'un modèle de substitution pour NEPTUNE_CFD, étape nécessaire car chaque calcul est réalisé en environ 2 heures de temps utilisateur avec 30 processeurs. Nous avons donc établi deux plans d'expériences numériques d'environ 2000 simulations chacun, le premier servant à la construction du modèle de substitution et le second à sa validation. Puisque la sortie du code est multidimensionnelle, il est nécessaire de recourir à la technique d'ACP mentionnée

précédemment. Deux modèles de substitution sont en fait créés : l'un pour la prédiction du taux de vide, et le second pour le diamètre de bulle. L'erreur de ces modèles est ensuite calculée sur le plan de validation, et est inférieure à 2% dans les deux cas. Nous considérons que cette valeur est négligeable devant les incertitudes expérimentale et de modèle présentes dans les simulations. A la suite de cette analyse, les modèles de substitution sont utilisés directement dans la fonction de vraisemblance en remplacement des évaluations de NEPTUNE_CFD.

Pour démontrer l'utilité de prendre en compte l'erreur de modèle dans cet exemple, nous réalisons une première calibration où sa valeur est imposée à 0. Le problème se ramène alors à une recherche des paramètres calibrés qui produisent le meilleur *fit* aux données expérimentales. Nous illustrons qu'il n'existe pas, dans le domaine de variation considéré, de valeur qui produise des prédictions satisfaisantes de deux quantités d'intérêt en même temps : le *fit* au taux de vide est privilégié, conduisant à de mauvaises prédictions du diamètre. De plus, les intervalles de crédibilité attachés aux prédictions sont extrêmement restreints et non représentatifs de la véritable incertitude.

Une calibration préliminaire est réalisée en utilisant les observations de taux de vide uniquement. Comme dans le chapitre précédent, les méthodes FMP, KOH et Bayes sont comparées. La méthode KOH fournit des postérieures à variance diminuée par rapport aux deux autres méthodes, avec intervalles de crédibilité restreints. Cela est défavorable lorsque les prédictions sur le diamètre de bulle sont tracées, illustrant l'importance de la conservativité d'une méthode lors de la prédiction de quantités d'intérêt non observées.

L'ensemble des observations est inclus dans la seconde calibration, prenant en compte le taux de vide et le diamètre. Les trois méthodes de calibration fournissent alors des résultats très proches. L'inclusion d'erreur de modèle a pour effet de sélectionner des prédictions de modèle qui ont la même forme que les observations, au lieu de minimiser le *fit*. Par exemple, pour le diamètre de bulle, les prédictions choisies gomment la forte décroissance au centre du canal qui était non physique. Malgré cet effort supplémentaire, les prédictions du modèle augmentées d'intervalles de crédibilité ne sont pas suffisantes pour recouvrir la totalité des observations, ce qui est la marque d'une erreur de modèle trop importante dans cette application.

Nous avons réalisé une étude de sensibilité au noyau d'erreur de modèle sélectionné : en faisant varier progressivement la fonction de covariance d'une structure faible (Diagonale, Matern 1/2) à une structure forte (Matern 5/2, Squared Exponential), nous montrons que la forme des prédictions est de plus en plus valorisée par rapport au *fit* aux observations. Cela fournit des pistes pour choisir un noyau adapté selon l'application.

Les résultats de cette étude de calibration suggèrent que le flux chauffant mesuré est au moins 5% inférieur à la véritable valeur de l'expérience, que le terme source de fragmentation est sous-estimé et que la coalescence et le transfert de masse sont surestimés. Ces résultats servent d'indications générales pour l'amélioration des modèles utilisés pour le transport de l'aire interfaciale. Deux extensions possibles de cette étude sont envisageables : la première est de considérer des modifications plus profondes des

modèles (ex : modification de la coalescence à la paroi du tube), pour étudier de nouvelles formes de modèles. La seconde est l'inclusion de modèles différents à calibrer (ex : forces, condensation), si ils sont soupçonnés de pouvoir améliorer les prédictions. Ces deux pistes sont à explorer pour permettre de reproduire de manière satisfaisante les observations du cas A6 en utilisant NEPTUNE_CFD sans terme correctif.

Chapter 1

Model error in CFD simulations

Contents

1.1 Scientific models, error, and inductivism

A scientific model is a simplified, abstract representation of the world. Trying to predict a natural phenomenon using a model is like trying to cook your grandmother's lasagna by following a recipe: sometimes you come close to it, but there's always a little something that is missing. But what went wrong? You do not know; otherwise, you would have already corrected it. So the best way to improve your recipe is to try again and refine it until it suits your taste. The scientific method relies on the following steps: observe a phenomenon, devise a theory or a model, run an experiment, and compare the obtained data to what your model predicted. If they match, it's great! Your model can predict the phenomenon and probably be used for similar phenomena. If they don't match, even better! It is the chance to go back to your books and propose a new model that will be more accurate.

Can you be certain that the model will still predict other experiments correctly? You cannot, says the falsifiability theory of Popper, as a theory or model can only be proved false and never be proved true. Sure, argue back the inductivists, but knowing that the model performs well on a wide range of experiments gives scientific credit to it and increases our belief that it will be as good on the next one. The idea of induction is familiar to anybody, even before any mathematical demonstration: taking the famous example of Hume, imagine a man who, every day, wakes up and sees the Sunrise in the East and who does not know the rotation of Earth around its axis. He can not be certain that, the day after, the Sun will rise in the East and not in the West. Despite

that, the more he makes the same observation, the more he will hold this belief to be true. But how does his belief increase every day? Does it increase by the same value every observation he makes? Is it even possible to calculate such a quantity? The answer to these questions lies in representing inductive logic with Bayesian Probability Theory. It provides a tool to reason, to evaluate the degree of veracity of logical propositions, and ultimately the degree of validity of scientific models. The Bayesian interpretation of probability is the one we adopt in this work and is described in the next section.

A potentially significant source of error in a model is that the assumptions made to establish it were wrong or too simplistic. It is called model error, or epistemic uncertainty, stemming from our lack of knowledge about the described phenomenon. It is impossible to create a model without assumptions, as before we describe something we require to be able to think about it, which is done by first conceptualizing the object, and by doing so, we can only create a view, capture a part of the true nature of the object. When I refer to the cup of tea placed on my desk, I could always go further in the description: the material it is made of, its patterns, etc. But the description needs to stop somewhere, finding a balance between simplicity and accuracy. The same goes for scientific models: they vary in degree of precision, and in the best scenario, if one model is insufficient, we can formulate a more accurate one. Take the example of trying to calculate the trajectory of a football after a free-kick from a player: a naive model would neglect air friction, and from the law of gravity, calculate a trajectory in the shape of a parabola. A few kicks suffice to determine that this model was wrong, so a new model is formulated that considers air friction. This new model performs better, but it is still not exact. To advance in the description, one could then take into account the deformation of the ball, and so on. By doing so, we always get closer to the truth without ever reaching it.

This fact should not leave the reader to give up all hope about science because we can never achieve a perfect description of the world. A model is not an end in itself, and the aim of the scientist is not to predict with extreme precision the ball trajectory; it is simply to score a goal. Thus, the estimation of model error is always goal-oriented: to which degree of precision is the solution required to be? The methods that we develop are universal in the sense that the same statistical methodology can be applied to all domains related to "hard" sciences: mechanical engineering, physics, chemistry, economics or epidemiology, as it requires only a model describing a phenomenon and experimental data obtained from this phenomenon. Paradoxically, they are also application-specific: as illustrated many times below, probabilistic methods can work out the answer to a given question, but a relevant question can only be asked by someone who has a solid knowledge of the domain. Besides, what to do with the answer, which decision to take when results are obtained, belong to the specialist. Thus, in this thesis, the methods are general and their application particular. In the sections that follow, we present briefly where error estimation techniques find their place in the context of two-phase CFD simulations for nuclear applications.

1.2 Bayesian considerations about scientific models

1.2.1 The logic interpretation of Bayesian probabilities

The Bayesian paradigm on probabilities has seen a tremendous increase in popularity since 2000 and is predominant in numerous fields such as information theory, statistical inference, image processing, or machine learning. Note that all these fields benefited from (or were created by) the exponential increase of computational power that came in this period, which is at the origin of numerous successful applications of Bayesian methods. But the Bayesian paradigm is not young: it originated from the article from Thomas Bayes *An Essay towards solving a Problem in the Doctrine of Chances* [Bayes and Price, 1763] in 1763, and an early adopter is Laplace in his *Essai Philosophique sur les Probabilités* [Laplace, 1814], which was highly criticized by his peers as a misuse of probabilistic calculations, applied to astronomy. Another reason why the "Bayesian revolution" came so late is that it goes against the well-established frequentist school of statistics, or "orthodox statistics", which dominated the view of statistical inference since its creation by R. A. Fisher in his influential treaty *Statistical Methods for Research Workers* [Fisher, 1925].

Under the Bayesian viewpoint, the probability of a logic proposition A, noted P(A), is interpreted as the degree of belief that A is true, rated from 0 (certain to be false) to 1 (certain to be true). Immediately the question comes to mind: why should probabilities be subjective, dependent on an external observer? A coin toss does not care about what an external observer thinks of it, so not all Bayesian probabilities are wrong most of the time and correctly assigned only by chance? To complete the definition of a probability in the Bayesian sense, we say that it is *the degree of belief that a rational thinker assigns to a proposition A, given his background information.* Thus we see that the notion of knowledge is deeply grounded in Bayesian probability theory. A person whose only information is that a coin toss is about to occur and knows nothing about the coin will assign a probability of 50% for it to come up with tails due to the lack of discerning information. Another person who knows that the coin is biased will assign a different probability to the same event. Thus both are correct as they assigned a probability based on the information they had at hand. The notion of "rational thinking" is defined by Cox [Cox, 1946] and consists of five *desiderata* that any legitimate reasoning should follow. For example, the consistency desideratum states that if there are multiple paths one can follow to find a result, the final result should be the same in all of them. These rules lead to one simple and elegant formula that dictates how one should assign probabilities to events based on the available information, which is Bayes' Theorem.

Bayes' Theorem is used to reason about conditional probabilities, which are noted for example P(A|B), read "the degree of belief that proposition A is true, given that proposition B is true". The joint probability of two events, noted $P(A \cap B)$, is the degree of belief that both are true. These probabilities are linked by the relation:

$$P(A \cap B) = P(A|B) * P(B) = P(B|A) * P(A), \tag{1.1}$$

and Bayes' Theorem is simply a rephrasing of the previous equation:

$$P(A|B) = \frac{P(A) * P(B|A)}{P(B)}. \tag{1.2}$$

In this single equation lies a lot of our intuition: and when both appear in contradiction, we always find, after careful examination, our intuition fooled us. Here are a few examples:

- If the knowledge of A does not modify our beliefs about B, we have $P(B|A) = P(B)$, so following Bayes' Theorem, $P(A|B) = P(A)$. Thus the knowledge of B will not modify our beliefs about A either. Following eq. (1.1), we obtain $P(A \cap B) = P(A) * P(B)$: it is a nice interpretation of the notion of independence between two events.
- Suppose that, if A is true, B is true. We have $P(B|A) = 1$, and so $P(A|B) = P(A)/P(B) \geq P(A)$, thus, observing that B is true will increase the degree of belief about A. It will be a substantial increase if B was an unlikely observation at first ($P(B) \approx 0$), but if B was already likely to be observed ($P(B) \approx 1$), the increase will not be substantial.

Many more illustrations can be found in the illuminating book [Jaynes and Bretthorst, 2003], where the authors defended the analogy between inductive reasoning and Bayesian Probability Theory vehemently.

1.2.2 Fundamentals of Bayesian Inference

Now let us approach the subject of Bayesian Inference, which is simply the application of Bayes' Theorem when dealing with scientific models or hypotheses, and experimental data. Traditionnally we note H an hypothesis, and D the data. Bayes' Theorem applies like so:

$$P(H|D) = \frac{P(H) * P(D|H)}{P(D)}. \tag{1.3}$$

Thus, the quantity of interest is $P(H|D)$, which is the degree of belief that assumption H is correct, knowing that the data D was observed. It is called the *posterior probability* of H. Note that it depends on $P(D|H)$, which is the probability of the data D occurring if hypothesis H is correct. A central consideration is that it also depends on $P(H)$, the *prior probability* of H, which is the degree of belief we had that H is true before observing the data D. It also depends on the general probability of observing the data $P(D)$.

In frequentist statistics, the quantity of interest is $P(D|H)$: start by assuming the model is true, then work out all possible data that could have occurred from the experiment that you ran, and compute the probability of obtaining this specific dataset. It is also called the p-value. Choose a level of probability arbitrarily, say 5%, and if the p-value is below that threshold, consider that the assumption H is rejected as the data was unlikely to have come out of it. In his essay, Bayes completely reversed this focus, as he considered that there was no use to consider hypothetical datasets that did not occur: what is known for certain is that the data D was observed, and what is uncertain

is the validity of the hypothesis H. See that the two points of view are linked by a simple formula, which explains that in a large variety of cases, both approaches will provide similar or identical results.

The dependence of the final result on the prior probability P(H) is likely the most criticized point in Bayesian theory: how can we evaluate the validity of an assumption H before seeing the data? It is done by considering the scientific credit granted to H: if one theory has already been successful in many applications, we can expect it to work out the same. On the contrary, a new theory will be given a low prior probability as it carries little weight. This case is where the expert knowledge of the topic comes into play, as it is required to assess the prior probabilities accurately. One could argue that if we had picked a different expert, he could have chosen entirely different prior probabilities, so Bayesian Inference is entirely subjective, whereas a scientific method should be objective. To this, we reply that all scientific studies are subjective and bear the mark of their author, be it on the assumptions taken, the models used, or the way the data is obtained, and so any scientific conclusions depend on prior assumptions. Bayesian Inference has the decency to explicitly make them appear in the form of prior probabilities, and in case you disagree with them, you can formulate new ones, apply Bayes' Theorem in precisely the same way, and see how the conclusions differ.

Bayesian methods can evaluate the degree of belief in a model, which is relevant mainly when two or more models are considered. Thus, it is a natural tool for model comparison, which can take the form of model improvement tests: compare two nested models, one being more accurate than the other, and see if the gain in performance is worth the added cost. Model-checking can also be performed by drawing posterior samples of the quantity of interest and seeing if they are physically acceptable. Many textbooks in Bayesian Inference propose applications of these concepts with statistical models (example of a question: do the observations come from a Gaussian Law or a Poisson Law?), see [Robert, 2007; MacKay, 2003]. To see it applied to physical models, the keywords are "Bayesian Inverse problems" [Stuart, 2010; Dashti and Stuart, 2017].

1.3 Industrial context: CFD simulations with nuclear applications

1.3.1 System codes in thermal-hydraulics

Numerical simulation is an essential tool to study the variety of scenarios that occur in nuclear plants. Here the focus is put on TH, the study of fluid flows that involve consequent heat transfers and thermodynamic phenomena such as phase change. These events are fully multiscale: taking the example of boiling, at the finest level, one could simulate each bubble for an accurate description. At the intermediate scale, the flow through one component, say, the nuclear core has to be considered for peak temperatures or heat transfer efficiency studies. At the largest scale, the global behaviour of the power plant must be represented via the interaction of its components like pumps, the core, or steam generators. Due to the computational cost, it is out of the question to simulate all these scales at once. Thus specific simulation tools were designed for each use. So-

called system codes like CATHARE[1], TRACE or RELAP5[2] are used for large-scale simulation of power plants. At the intermediate scale, codes are dedicated to specific components; such examples at CEA are the codes GENEPI+ for steam generators and FLICA5 for the core. Fine-scale simulations are realized with CFD solvers, of which a large selection exists depending on the application, some examples are TRIO_CFD[3], NEPTUNE_CFD[4], or Ansys Fluent[5].

For example, with the system code CATHARE, the hydraulics are represented using six equations [Bestion, 1990] which correspond to conservation equations for the mass, momentum and energy of the liquid and gas phase. Each of these equations involves *closure laws*, which represent the exchange terms of mass, momentum and energy between the two phases and between the phases and the walls. These laws require heavy physical modelling, as they need to bear on averaged quantities (due to the very large simulation scale) while being general enough to represent a variety of physical phenomena. Thus, modelling errors are prone to occur, so there is a need for additional proofs of the adequacy of the numerical results, in the form of V&V and UQ processes, detailed in the next section.

Note that the focus in this work is not put on the system codes themselves, but rather the fine-scale CFD simulations that can serve as reference data. These can range from the simulation of a swarm of bubbles to mixing cold and hot fluids in a T-section of a pipe. When simulating extremely local phenomena, like the bubble swarm, it is possible to create a fine mesh to capture all the physics. It requires approximately ten cells per bubble diameter, and then the Navier-Stokes equations are directly solved without resorting to closure laws: this is the DNS. Such a feat is already too expensive for the simulation of the T-junction: thus, in CFD, averaged approaches are used, notably for turbulence models, using LES for mid-sized cells, and RANS for large-sized cells. In multiphase CFD, which will be the topic of application of our work, averaged formulations are the reference, working with balance equations for the phases and exchange terms that are subject to model error.

1.3.2 Verification, Validation and Uncertainty Quantification

Before the results obtained with a computer code can be accepted as scientific elements in studies related to the nuclear field, the code must go through Verification and Validation steps. Each of these steps is dedicated to one type of uncertainty: in the Verification step, the uncertainties of numerical nature are examined. These involve typically mesh refinement studies or code bugs. In the Validation step, the simulation results are compared to experimental data representing the physical phenomena involved.

From early uses of system codes in the 1950s to now, V&V steps have been used as a guarantee of simulation accuracy. Evolving with the maturity of simulation tech-

[1] https://cathare.cea.fr/
[2] https://www.nrc.gov/about-nrc/regulatory/research/safetycodes.html
[3] https://triocfd.cea.fr/
[4] https://www.code-saturne.org/cms/web/NEPTUNECFD
[5] https://www.ansys.com/fr-fr/products/fluids/ansys-fluent

niques and the surge of numerical power that enables more in-depth studies, they are provided in the form of guidelines, generally field-specific. Modern references for CFD are [AIAA, 1998], and [ASME, 2009], established respectively by the American Institute of Aeronautics and Astronautics and the American Society of Mechanical Engineers. UQ constitutes the third step for the accuracy assessment of codes, much younger than V&V, and comes in complement of them. It aims to quantify the different sources of error in the simulations and evaluate the variability of numerical predictions that results from them. UQ studies have been pushed forward as an effort to promote the BEPU methodology [D'Auria et al., 2012] for reactor transient simulations and safety analysis. This methodology encourages the use of numerical simulations attached with uncertainties to be compared with the risk margins present in the domain. This concept opposes the traditional conservativism employed in nuclear safety, which encourages heavy safety margins when dimensioning a system, possibly leading to a hefty additional construction cost. The international projects PREMIUM [CSNI, 2016] and SAPIUM [Baccou et al., 2020] aimed to formulate guidelines for UQ techniques in the nuclear context.

In thermal hydraulics, a popular UQ technique is the fUQ, in which the uncertainties on the input parameters are run through the computer model to evaluate the variability on the quantity of interest. This is implemented for example in the CSAU method [Boyack et al., 1990; Zhao and Mousseau, 2012] or the GRS method [Glaeser, 2008]. Another equally important family is the ensemble of iUQ methods, which aim to determine the uncertainty on model input parameters by using experimental data. These are often referred to as Inverse problems or Backwards problems. A survey of iUQ methods applied to system codes is proposed by [Wu, Xie, et al., 2021], who suggests a classification into three categories: frequentist methods which look for "best-fit" parameters (e.g. CIRCE [De Crecy and Bazin, 2004], later extended to Bayesian [Damblin and Gaillard, 2019]), Bayesian methods which treat input parameters as random (e.g. Modular Bayesian Approach [Wu, Shirvan, et al., 2019]), and empirical methods which exploit samples of the input parameters to fulfil the desired agreement to data (e.g. DIPE [Joucla and Probst, 2008]).

The iUQ methods, similarly to the Validation techniques, are largely data-based. Consequently, the choice of experiments to be used is central. Due to the multiphysics and multiscale character of phenomena in TH, SET are privileged to serve as a basis for the establishment of closure laws or assessing the uncertainty on existing ones. These tests aim to reproduce individual physical phenomena by limiting their interactions with others. Generally the experiments are designed with a reduced geometry compared to the true application, so the scale factor is an additional source of uncertainty, which is not studied in this work. For a more detailed discussion about scale factor uncertainty, see [D'Auria et al., 2012; Ivanov et al., 2018]. In addition to the SETs, the IET combine multiple phenomena and are also used for validation purposes. However, due to the complexity and the size of the systems involved, IET validation experiments are few, and voluntarily inducing incidents on existing power plants to obtain measurements is not common practice for evident reasons.

CFD simulations, in this context, have the potential to serve as "numerical experi-

ments" to provide reference data for uncertainty studies. Numerical simulations do not carry the same weight as real experiments, but their use is constantly increasing due to the growing maturity of the tools and the relatively low engaged costs. Thus, it is necessary to perform UQ studies on the CFD simulations themselves and improve their reliability for further uses. To the best of our knowledge, there are only a handful of studies in the scientific literature focused on the determination of model error in CFD simulations, even less in the nuclear context.

1.4 Challenges and contributions

Objectives and challenges

As was discussed in the previous section, TH simulations in the context of nuclear engineering involve multiscale, multiphysics coupled problems with highly non-linear interactions. Model uncertainties cannot be avoided due to the complexity of the phenomena involved and the lack of experimental data required to formulate new models and validate the existing ones. Thus, it is necessary to estimate the uncertainties accurately in a robust manner and show their impact on model predictions. A proposed statistical method for error estimation should answer the following objectives:

- **Quantify modelling uncertainties** that apply to the models involved (ex: phase interactions, heat fluxes, or boiling) and evaluate their impact on predictions. This problem is challenging, especially when the initial model strongly under- or over-predicts some quantities and is sensitive to its input parameters.
- **Quantify experimental uncertainties** that often play a decisive role in UQ studies. They must be carefully considered, as it is delicate to formulate a relevant statistical model to include them in most cases.
- **Find relevant values for model parameters**, with the inclusion of the two sources of uncertainty below. An accurate determination of coefficients is crucial for determining the range of applicability of models. If the uncertainties are not correctly determined at this step, the determined parameter values can be erroneous and falsely confidence-inducing.
- **Perform the above on a large experimental database**. This step requires the method to be scalable and inexpensive. It is crucial for model validation, as only by including data in a wide experimental range can a model be considered suitable for use in general conditions. By conducting studies with only a few experimental cases, we risk obtaining conclusions that are valid only in a limited domain.
- **Identify the possible improvements of a physical model**. If the model uncertainties are estimated at too high of a level for the considered application, the UQ framework should provide hints to show where further modelling efforts are necessary.

The context of the application is also the source of additional challenges:

- **Work on models with a large number of parameters**. As multiphase CFD makes use of tens of models or experimental correlations for physical phenomena, the total number of parameters can be of the order of hundreds or more, depending on the desired depth of modifications. Often UQ techniques can not deal with so many due to the *curse of dimensionality*, so this requires a preliminary selection, either from statistical or physical considerations.
- **Work with solvers that are expensive to evaluate**. Two-phase flows simulations require intensive calculations, usually done on supercomputers. Our applications are relatively "cheap", each calculation taking 2 hours of user time on 30 cores. Statistical techniques require many calculations for better precision. Thus, a compromise is to be found.
- **Deal with strong non-linear behaviour of code response**. Due to the complexity of the numerical solver, some strong non-linearities with respect to the model parameters are to be expected. This feature can only be appreciated adequately with sufficient code evaluations, with the issues mentioned in the previous points.
- **Eliminate other potential sources of errors**. In unsteady simulations, two sources of numerical error are considered: spatial discretization and time convergence. Both can induce a variation of the code output that could be mistaken with model error. Thus, it is necessary to ensure that these sources are properly controlled.

Contributions:

In this thesis, we offer two kinds of contributions: the first one is theoretical and numerical, in the form of a new calibration framework and associated algorithms. The second kind is applicative, demonstrating the developed framework on two concrete examples. The main theoretical and numerical contributions are briefly summed up as follows:

- **A new estimator for model error** that is theoretically grounded. It was chosen to be less restrictive than what is proposed in the literature to accommodate a broader range of shapes of model predictions. We propose a new definition of model error which solves the identifiability issue between model uncertainties and parameter uncertainties.
- **The integration of this estimator into a calibration framework** which allows searching for plausible parameter values by considering all sources of uncertainties. We reduce the computational cost by introducing the FMP approximation based on optimal hyperparameters. The theoretical properties of the method are studied, notably in three asymptotic regimes where the number of observations tends to infinity.
- **A surrogate model-based acceleration technique**, to reduce the cost of the entire calibration. Optimal hyperparameters are represented with surrogate models

using an adaptive algorithm that places training points in the posterior plausible range, consistently with the whole calibration framework.

The two applications are described below:

- **A heat flux partitioning model** that is cheap to evaluate, in which we calibrate an experimental correlation and the contact angle between water and a steel surface. It is done using the heat flux values in the Kennel database of 13 boiling experiments with a separate model discrepancy term for each, for a total of 26 unknown hyperparameters, which is consequent for this type of study. In this situation where no single value of the parameters can reproduce all the experiments accurately, it is essential to consider a model error. We study both individual and global calibrations. The FMP method proved to be the most accurate, robust and scalable method.
- **NEPTUNE_CFD two-phase flow solver**, in which the uncertainty in the interfacial area transport model is evaluated. It is done using data from one DEBORA bubble column experiment, including the experimental uncertainty on the heating flux at the walls. We work with a finite budget of costly solver evaluations. Thus, it is needed to resort to surrogate modelling and numerical Design of Experiments techniques. Two quantities of interest are dealt with, the radial profiles of void fraction and bubble average diameter. Calibrations are performed on the different quantities, then with both. In-depth physical insights, coupled with preliminary simulation studies, are required to set up a proper calibration framework. We also study the impact of prior choice for the model discrepancy, including the calibration with no model error at all. The FMP method also proves to be an excellent compromise between accuracy and computational cost.

1.5 Outline

Chapter 2: In this chapter, we propose a literature review of the Bayesian calibration framework, applied to computer codes with consideration of model error. The popularity of this procedure started with the seminal work [Kennedy and O'Hagan, 2001a], which is presented in detail. We propose to look at multiple methods in the literature under the scope of how they estimate model discrepancy hyperparameters. We also discuss the identifiability issue that comes from the model error definition of Kennedy and O'Hagan. Then we review methods for solving practical questions when one wants to try his hand at calibration, namely the construction of a surrogate model for the code or the choice of kernel.

Chapter 3: We expose our theoretical contributions to the calibration with model error are presented. We propose a new definition of model error, the FMP approximation and its theoretical grounds. We show that when the posterior distribution is a mixture of Gaussians, the FMP method outperforms the approach of Kennedy and O'Hagan and prevents the false confidence effect. We prove regularity properties of the quantities in-

volved that will serve as a guarantee for the good working of the method. We propose a calibration study on a numerical example when the number of experimental observations tends to infinity under three asymptotic frameworks. This procedure allows determining if, in practice, the Gaussian approximation is relevant.

Chapter 4: The chapter contains numerical methods, novel or already known, that are relevant to calibration and will be used in the applications chapters. Classical techniques involve sampling the posterior distributions using MCMCs and surrogate building of a multi-valued computer code by PCA. We present these algorithms and their specificities when applied in our calibration framework. We introduce a resampling algorithm to improve the quality of the FMP sample. We also present the surrogate modelling strategy inside the calibration methodology and an adaptive algorithm to choose surrogate training points in the posterior plausible region. These techniques are illustrated in elementary numerical examples.

Chapter 5: This chapter presents a calibration study of a heat flux partitioning model using multiple experimental configurations. The goal is to improve an existing experimental correlation by considering new data from a boiling experiment. The model predictions on the experiments are also enhanced as they become more accurate and augmented with credibility intervals. The experimental uncertainties bear on the input quantities of the model instead of the output, which requires special effort. We study the calibration results of each experiment individually and on the entire set. The FMP calibration is compared with the KOH calibration in both cases, and the reference solution is attainable when considering a subgroup of well-behaving experiments. This fact illustrates the robust character of our method and its applicability.

Chapter 6: Here, we present the second application of our calibration technique on a CFD solver that is used to reproduce a two-phase flow in a vertical pipe. After proposing physically justified leads on improving the models involved in the calculation, we show how to choose the relevant leads, arriving to the choice of model modification proposed. Then, our calibration techniques apply to two quantities of interest: the void fraction and bubble diameter. We first consider them separately, then in a joint manner. Once again, the FMP, KOH and reference calibration techniques are applied, showing the robust character of our technique. We show that performing calibration without consideration of model error is misleading and investigate the impact of the choice of kernel for the model discrepancy.

Chapter 7: We summarize the results obtained in this thesis and propose research leads for further progress in the topic.

Chapter 2

Calibration of computer codes

Contents

In this chapter, we propose a specialized literature review about the Bayesian Calibration framework applied to computer codes. This topic was arguably instigated by the KOH framework, which is presented in detail. Then, we present the more recent variants of the framework under the angle of how the hyperparameter estimation for the model discrepancy term is performed. Then, we can discuss the identifiability issue that we aim to correct. The third section of the chapter proposes a literature review about practical questions that arise when applying this framework.
In this chapter and the ones that follow, the reader is assumed to be familiar with the concepts of random processes and Gaussian Process Regression, in particular with noisy observations.

2.1 Brief history of calibration

Computer codes typically involve parameters whose value is not known with exactitude. Some of these parameters hold a physical meaning, and some are usually considered tuning parameters that can be adjusted. An intuitive way to choose the values of these parameters is to navigate by "trial and error", where various configurations of parameters are tried to see which one fits the experimental observations the best. Calibration techniques emerged in the 80s when it became clear that there was a need for more principled estimation methods as the role of computer codes became prominent in scientific studies. Statistical methods for parameter estimation were available, but they relied on

explicit probability models and did not apply to computer codes, of which one run could take multiple days. Thus, significant progress was made in this field when practitioners started *emulating* computer codes, i.e. building simple probability models that were cheap to evaluate, making a thorough exploration of the parameter space possible. Such early efforts can be found in [Miller and Frenklach, 1983] with application to chemical kinetics, [Gregory and Smith, 1990] in econometrics, and [Capodieci et al., 1998] in electrical engineering. The simple probabilistic model that emulates the computer code behaviour is referred to as the surrogate model.

Long before computer codes, scientists developed regression techniques, which, starting from experimental data, allowed to build statistical models to emulate the behaviour of the real system. Thus, the much more recent issue of building a surrogate model using data obtained from computer runs naturally took inspiration from the regression techniques, giving birth to the name "computer experiments" to designate the simulated data. A question that is precious to both domains is the manner of choosing the best locations to run experiments. This is referred to as the DoE [Fisher, 1925] in the first case, and the nDoE [Sacks et al., 1989] in the second, of which first efforts can be traced back to 1979 [Mckay et al., 1979]. Note that the names refer both to a research field and the collection of experimental points obtained. Still, substantive differences remain between the two fields: the widely used methodology of response surfaces for experimental data [Box and Wilson, 1951] is not well suited for computer experiments because the latter are not subject to random error [Currin, 1988].

One of the most popular classes of surrogate models is Gaussian Processes (GPs), which were derived from the kriging techniques originating in the years 1950 [Matheron, 1963], using experimental data. The ingenious idea of kriging is that the data under study is considered a realization of a random process, which allowed for a full statistical framework for data assimilation, prediction and uncertainty evaluation. This concept will be discussed in more detail in section 2.3.1. The kriging techniques were applied to a large extent to computer experiments for systematic research of best parameters values. Two influential works that offer a mature viewpoint on these applications before the years 2000 are [Sacks et al., 1989] for experimental design methods, and [Cox et al., 2001] for an integral calibration framework.

Among these applications, Bayesian methods emerged for calibration. In the generalized likelihood uncertainty estimation technique [Beven and Freer, 2001], originally applied to hydrological models, a Monte Carlo sample of parameters is pushed through the computer model and weighted with likelihood factors, effectively making Bayesian predictions. In the works of [Craig et al., 2001] the problem of history matching is addressed, which amounts to calibration on a time series. A Bayesian strategy is employed, heavily based on experts' prior beliefs. Instead of performing full posterior calculations, only expectancies and variances are computed according to the Bayes linear framework [Goldstein and O'Hagan, 1996]. This trick helps to reduce the added cost of calibration using time series instead of point data. In [Raftery et al., 1995], calibration is performed by including prior information on both model parameters and model outputs, which they name the *Bayesian synthesis* technique. This method was later refined into *Bayesian*

melding [Poole and Raftery, 2000]. Bayesian interpretation of kriging methods for calibration and design of numerical experiments appeared notably in [Currin, 1988] and [Morris et al., 1993].

The article [Kennedy and O'Hagan, 2001a] offered a new perspective on Bayesian calibration. The author's idea was to formulate a framework in which all uncertainties that plague the computer model predictions would be taken into account. For instance, in previous calibration frameworks such as in [Cox et al., 2001], the model parameters would be set to their best fit value, leading to an underestimation of the parameter uncertainty. The central idea in their work is to introduce an additional error term in the calibration equation to represent the fact that a computer code cannot, in general, exactly reproduce experimental observations. This term, named *model discrepancy* or *model bias*, can also be informed using expert beliefs about how the computer code is biased.

2.2 The Bayesian framework of Kennedy and O'Hagan

2.2.1 Multiple sources of uncertainties

We present a simple classification of uncertainties as defined in [Kennedy and O'Hagan, 2001a]:

- **Parameter uncertainty**: this is due to the fact that the values of parameters in the model are not known with exactitude, which is a source of additional variability in the model predictions.
- **Model inadequacy**: or model error. Even when using the best parameter values, an irreducible discrepancy remains between the model predictions and the observations due to a lack of knowledge about the described phenomenon.
- **Measurement uncertainty**: it is caused by non-perfect measure instruments, source of an extra uncertainty on the data points. This aspect is central in the Bayesian framework since the data is the information source for learning about the model.
- **Code uncertainty**: it is present when the computer code is replaced with a surrogate model for cost efficiency. It also induces an approximation error of the code results, which hinders our ability to learn about the model parameters.

These uncertainties are all taken into account in their Bayesian framework. Note that additional uncertainties could also be considered. For example, they mention the *residual variability* that emerges from uncontrolled or unknown factors in experiments that lead to poor reproducibility and an additional uncertainty on the data points. This variability is dealt with by averaging the data points over multiple realizations of the experiments. One could also introduce more refined types of model inadequacy. As formulated in KOH, it is the difference between the result of the computer model and the true process. However, the result of the calculation is only the numerical approximation of the solution of a physical model. The Bayesian framework offers a natural way to include this numerical

error, yet it is not considered in KOH's work because other techniques such as code validation are generally used to evaluate this source of uncertainty.

2.2.2 Problem formulation

Consider an observable quantity measured through multiple runs of an experiment with variable experimental conditions. The experimental conditions are noted $\mathbf{x} \in \mathbb{X}$, the observed process $y_{\text{obs}}(\mathbf{x})$ and the true process $y(\mathbf{x})$. For simplicity of exposition, we deal only with scalar observables, and we suppose that $y(\cdot)$ is not subject to residual variability. The relationship between the two processes writes :

$$y_{\text{obs}}(\mathbf{x}) - y(\mathbf{x}) = \varepsilon(\mathbf{x}), \tag{2.1}$$

where $\varepsilon(\mathbf{x})$ is a random process that represents measurement error.

We work with a computer code f that takes as inputs the experimental conditions and model parameters $\boldsymbol{\theta} \in \Theta$ and outputs a prediction of the observable $f(\mathbf{x}, \boldsymbol{\theta})$. We assume that there exists a best value of model parameters $\boldsymbol{\theta}^*$, and so the model discrepancy $z(\cdot)$ is defined as :

$$z(\mathbf{x}) := y(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\theta}^*). \tag{2.2}$$

The calibration is performed using n observations $\mathbf{y}_{\text{obs}} = (y_{\text{obs},1}, \ldots, y_{\text{obs},n})^T$ obtained from experiments. The experimental conditions are summed up in the observed dataset $\mathrm{D}_2 = (\mathbf{x}_{\text{obs},1}, \ldots, \mathbf{x}_{\text{obs},n})$. We also have another dataset that consists of m computer code runs on the simulated dataset $\mathrm{D}_1 := ((\mathbf{x}_{\text{calc},1}, \boldsymbol{\theta}_{\text{calc},1}), \ldots, (\mathbf{x}_{\text{calc},m}, \boldsymbol{\theta}_{\text{calc},m}))$, and the results are stored in the vector $\mathbf{f}_{\text{calc}} = (f_{\text{calc},1}, \ldots, f_{\text{calc},m})^T$. Typically we dispose of many more computer runs than experimental runs, so that $m >> n$.

Combining equations (2.1) and (2.2), we obtain the calibration equation that specifies the link between the observed process and the computer model predictions :

$$y_{\text{obs}}(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}^*) + z(\mathbf{x}) + \varepsilon(\mathbf{x}). \tag{2.3}$$

The traditional assumption is to take gaussian measurement errors with mean zero, fixed variance σ^2_{mes}, and uncorrelated to each other. We can thus write $\varepsilon(\mathbf{x})$ as a Gaussian process :

$$\varepsilon(\mathbf{x}) | \sigma_{\text{mes}} \sim \mathrm{GP}(0, \sigma^2_{\text{mes}} \delta(\mathbf{x}, \mathbf{x}')),$$

where $\delta(\mathbf{x}, \mathbf{x}') = \mathbb{1}_{\mathbf{x}=\mathbf{x}'}$. We also suppose the following GP structure for the model discrepancy term :

$$z(\mathbf{x}) | \boldsymbol{\beta}_z, \boldsymbol{\psi}_z \sim \mathrm{GP}(\mathbf{h}_z(\mathbf{x})^T \boldsymbol{\beta}_z, c_{\boldsymbol{\psi}_z}(\mathbf{x}, \mathbf{x}')),$$

where $\mathbf{h}_z(\mathbf{x}) = (h_{z,1}(\mathbf{x}), \ldots, h_{z,p}(\mathbf{x}))^T$ is a vector of p user-specified regression functions, and $c_{\boldsymbol{\psi}_z}$ is the covariance function. The model discrepancy distribution is a function of location parameters $\boldsymbol{\beta}_z$ of size p, and covariance hyperparameters $\boldsymbol{\psi}_z$.

A third Gaussian process structure is assumed for the computer code:

$$f(\mathbf{x}, \boldsymbol{\theta}) | \boldsymbol{\beta}_f, \boldsymbol{\psi}_f \sim \mathrm{GP}(\mathbf{h}_f(\mathbf{x}, \boldsymbol{\theta})^T \boldsymbol{\beta}_f, c_{\boldsymbol{\psi}_f}((\mathbf{x}, \boldsymbol{\theta}), (\mathbf{x}', \boldsymbol{\theta}'))),$$

where $\mathbf{h}_f(\mathbf{x}, \boldsymbol{\theta}) = (h_{f,1}(\mathbf{x}, \boldsymbol{\theta}), \ldots, h_{f,p'}(\mathbf{x}, \boldsymbol{\theta}))^T$ are $p^{'}$ user-specified functions and $c_{\boldsymbol{\psi}_f}$ is the covariance. This distribution also depends on location parameters $\boldsymbol{\beta}_f$ and hyperparameters $\boldsymbol{\psi}_f$. We note $\boldsymbol{\beta} = (\boldsymbol{\beta}_f, \boldsymbol{\beta}_z)$ and $\boldsymbol{\psi} = (\boldsymbol{\psi}_f, \boldsymbol{\psi}_z)$.

Let $\mathbf{H}_f(\mathrm{D}_1)$ be the $m \times p^{'}$ tendency matrix with (i, j) coefficient $h_{f,j}(\mathbf{x}_{calc,i}, \boldsymbol{\theta}_{calc,i})$. With similar notations, $\mathbf{H}_z(\mathrm{D}_2)$ is the $n \times p$ matrix with (i, j) coefficient $h_{z,j}(\mathbf{x}_{\mathrm{obs},i})$. We also require the mean tendency matrix of the surrogate model evaluated on D_2, which can be evaluated provided that $\boldsymbol{\theta}^*$ is known: $\mathbf{H}_f(\mathrm{D}_2)$ is the $n \times p^{'}$ matrix with (i, j) coefficient $h_{f,j}(\mathbf{x}_{\mathrm{calc},i}, \boldsymbol{\theta}^*)$.

All three Gaussian processes are supposed to be independent a priori. We note $\boldsymbol{\beta} = (\boldsymbol{\beta}_f, \boldsymbol{\beta}_z)$, $\boldsymbol{\psi} = (\boldsymbol{\psi}_f, \boldsymbol{\psi}_z)$ and $\mathbb{I} = (\mathrm{D}_1, \mathrm{D}_2)$ the background information. Following our statistical model, the probability densities for $\mathbf{f}_{\mathrm{calc}}$ and $\mathbf{y}_{\mathrm{obs}}$ are gaussian, and we write their means and variances below. Note that, for clarity, we condition to all unknowns when expressing the means and variances of $\mathbf{y}_{\mathrm{obs}}$ and $\mathbf{f}_{\mathrm{calc}}$. By doing so, we condition to unnecessary information, but the unknowns that are relevant to each formula can easily be read on the right-hand side of the equations.

$$\mathbb{E}[\mathbf{f}_{\mathrm{calc}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \mathbf{H}_f(\mathrm{D}_1)\boldsymbol{\beta}_f, \tag{2.4}$$

$$\mathbb{E}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \mathbf{H}_f(\mathrm{D}_2)\boldsymbol{\beta}_f + \mathbf{H}_z(\mathrm{D}_2)\boldsymbol{\beta}_z. \tag{2.5}$$

Variances can be explicited by noting $\mathbf{V}_f(\mathrm{D}_1)$ the matrix with (i, j) element $c_f((\mathbf{x}_{calc,1}, \boldsymbol{\theta}_{calc,1}), (\mathbf{x}_{calc,j}, \boldsymbol{\theta}_{calc,j}))$, with similar definition for $\mathbf{V}_z(\mathrm{D}_2)$. $\mathbf{V}_f(\mathrm{D}_2(\boldsymbol{\theta}^*))$ has for (i, j) element $c_f((\mathbf{x}_{\mathrm{obs},i}, \boldsymbol{\theta}^*), (\mathbf{x}_{\mathrm{obs},j}, \boldsymbol{\theta}^*))$. The correlation matrix $\mathbf{C}_f(\mathrm{D}_1, \mathrm{D}_2)$ has for (i, j) element $c_f((\mathbf{x}_{\mathrm{calc},1}, \boldsymbol{\theta}_{\mathrm{calc},1}), (\mathbf{x}_{\mathrm{obs},j}, \boldsymbol{\theta}^*))$.

$$\mathrm{Var}[\mathbf{f}_{\mathrm{calc}}\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \mathbf{V}_f(\mathrm{D}_1), \tag{2.6}$$

$$\mathrm{Var}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \mathbf{V}_f(\mathrm{D}_2(\boldsymbol{\theta}^*)) + \mathbf{V}_z(\mathrm{D}_2) + \sigma^2_{\mathrm{mes}}\mathbf{I}_n. \tag{2.7}$$

$$\mathrm{Cov}[\mathbf{f}_{\mathrm{calc}}, \mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \mathbf{C}_f(\mathrm{D}_1, \mathrm{D}_2(\boldsymbol{\theta}^*)). \tag{2.8}$$

The full data vector $\mathbf{d} = (\mathbf{f}_{\mathrm{calc}}, \mathbf{y}_{\mathrm{obs}})$ thus follows a gaussian distribution with parameters given by:

$$\mathbb{E}[\mathbf{d}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \begin{pmatrix} \mathbb{E}[\mathbf{f}_{\mathrm{calc}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] \\ \mathbb{E}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}], \end{pmatrix} \tag{2.9}$$

and:

$$\mathrm{Var}[\mathbf{d}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] = \begin{pmatrix} \mathrm{Var}[\mathbf{f}_{\mathrm{calc}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] & \mathrm{Cov}[\mathbf{f}_{\mathrm{calc}}, \mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}]^T \\ \mathrm{Cov}[\mathbf{f}_{\mathrm{calc}}, \mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}] & \mathrm{Var}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}]. \end{pmatrix} \tag{2.10}$$

The application of Bayes' theorem requires to specify prior information for all unknowns. The two most common choices for location parameters $\boldsymbol{\beta}$ are uniform and normal. Here we consider the uniform prior $\mathrm{p}(\boldsymbol{\beta}) \propto 1$. Typical priors for hyperparameters

$\boldsymbol{\psi}_f$ and $\boldsymbol{\psi}_z$ depend on the chosen covariance structure and the meaning of each hyper-parameter. We will discuss typical covariance structures and priors in section 2.3.3. For the measurement error variance, the classical prior is Jeffrey's prior $\mathrm{p}(\sigma_{\mathrm{mes}}) \propto 1/\sigma^2_{\mathrm{mes}}$. For optimal model parameters, the prior is determined according to expert opinion on the problem. We also suppose that a priori, all unknowns are independent, so that in our case:

$$\mathrm{p}(\boldsymbol{\theta}^*, \boldsymbol{\psi}, \boldsymbol{\beta}, \sigma_{\mathrm{mes}}) = \mathrm{p}(\boldsymbol{\theta}^*)\mathrm{p}(\boldsymbol{\psi})\mathrm{p}(\sigma_{\mathrm{mes}}). \tag{2.11}$$

The full posterior distribution then writes:

$$\mathrm{p}(\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}|\mathbf{d}, \mathbb{I}) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\psi})\mathrm{p}(\sigma_{\mathrm{mes}})\mathrm{p}(\mathbf{d}|\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\mathrm{mes}}, \mathbb{I}). \tag{2.12}$$

The posterior distribution of (2.12) can be used to find out the new beliefs about the best model parameters, updated with the information contained in the observed data $\mathbf{y}_{\mathrm{obs}}$ and the simulated data $\mathbf{f}_{\mathrm{calc}}$. To find out specific results about a subset of model parameters, one has to integrate out all non-desired variables. A single value of model parameters can be made out by estimation as maximum a posteriori or posterior mean to communicate the calibration results for further studies. To get a more accurate result, one can add posterior confidence intervals to this value in the form of tolerance bounds [Bayarri et al., 2007], or even provide the full posterior information by handing out a sample of the posterior distribution, keeping in mind that lengthy samples will be required to reproduce high-dimensional distributions accurately.

Kennedy and O'Hagan suggest using an approximation of the posterior distribution by estimating hyperparameters and location parameters and plugging these estimates into the likelihood function, leading to an approximation of the true posterior distribution of model parameters. This approximation can be used in the next steps of the analysis to predict the true process $y(\mathbf{x})$ and uncertainty analysis of the computer model. We give a general idea of this method in the next section.

2.2.3 Modular approach and prediction

Modular approach

Note that, by calculating the posterior distribution, we solve three coupled sub-problems at the same time:

1. The construction of a surrogate model for f,
2. The estimation of a model discrepancy term z and measurement error σ_{mes},
3. The calibration of model parameters $\boldsymbol{\theta}^*$.

A joint resolution of these three sub-problems is not desirable for the following reasons:

First, it is numerically expensive to compute a full Bayesian solution since the summed dimensions of the unknowns in the problem $(\boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi})$ can reach tens or even hundreds. A common way of getting samples from the posterior distribution is to employ

sampling algorithms such as MCMC-based algorithms, which may show poor mixing in these cases due to the difficulty of exploring a high-dimensional space.

Second, the estimation of the surrogate model for f theoretically depends on the real-world observations $\mathbf{y}_{\text{obs}}$, as can be seen in equations (2.5) and (2.7). Conceptually, it can be understood that these observations are supposed to emerge from simulations of the computer model at $\boldsymbol{\theta}^*$, up to the model discrepancy. In practical applications, we would prefer to build the surrogate model using only the simulated data $\mathbf{f}_{\text{calc}}$.

Third, a well-known cofounding issue occurs between the parameter uncertainty and the model discrepancy when they are estimated together. Proposing a separate estimation for both is a strong approximation often performed in practice. We discuss these identifiability issues in more details in section 2.2.4.

These arguments led Kennedy and O'Hagan to perform a sequential resolution of these three sub-problems. This approach was theorized [Liu et al., 2009] and named *modular approach*. In this approach, it is advocated that complex, multi-level Bayesian problems should be decomposed into "modules" that are to be treated sequentially. One major argument is to alleviate the numerical cost, as mentioned previously. Additionally, it is argued that often it is one module that is problematic and that can potentially ruin the entire Bayesian analysis. In contrast, treating them separately and performing independent validation could help find the problematic module and improve its treatment. The previous Bayesian calibration problem can be decomposed into three modules:

1. Compute estimators $(\widehat{\boldsymbol{\beta}}_f, \widehat{\boldsymbol{\psi}}_f)$ of the hyperparameters of the computer code using the simulation data alone $\mathbf{f}_{\text{calc}}$.
2. Compute estimators $(\widehat{\boldsymbol{\beta}}_z, \widehat{\boldsymbol{\psi}}_z, \widehat{\sigma_{\text{mes}}})$ of the hyperparameters of the model discrepancy and measurement error using the observations $\mathbf{y}_{\text{obs}}$ and the previous estimates $(\widehat{\boldsymbol{\beta}}_f, \widehat{\boldsymbol{\psi}}_f)$.
3. Plug the hyperparameter estimates in the likelihood function and get samples from the posterior distribution of $\boldsymbol{\theta}^*$.

Note that in [Liu et al., 2009] the measurement error is estimated in a separate module that should be solved before the model discrepancy module. We found in our applications that including its estimation in the model discrepancy module did not pose significant problems since there is little cofounding between the two. This fact is especially true when there is a strong correlation structure in z so that the problem amounts to splitting error between a part with spatial correlation (model discrepancy) and without (measurement error).

Prediction

The calibration is sometimes viewed as an intermediate step to the prediction step, in which the real process is estimated, and the estimator can be used for interpolation or extrapolation tasks. Under the KOH framework, the real process is a realization of a GP conditionally to the model parameters, which provides a convenient way of accounting for the uncertainty in its prediction. The mean function of the true process $y(\cdot)$ is given

by:

$$\mathbb{E}[y(\mathbf{x}')|\mathbf{d}, \boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\text{mes}}, \mathbb{I}] = \mathbf{h}(\mathbf{x}', \boldsymbol{\theta}^*)^T\widehat{\boldsymbol{\beta}} + \mathbf{t}(\mathbf{x}', \boldsymbol{\theta}^*)^T\mathbf{V}_d(\boldsymbol{\theta}^*)^{-1}(\mathbf{d} - \mathbf{H}(\boldsymbol{\theta}^*)\widehat{\boldsymbol{\beta}}) \quad (2.13)$$

with

$$\mathbf{h}(\mathbf{x}', \boldsymbol{\theta}^*) = \begin{pmatrix} \mathbf{h}_f(\mathbf{x}', \boldsymbol{\theta}^*) \\ \mathbf{h}_z(\mathbf{x}') \end{pmatrix}, \quad \mathbf{t}(\mathbf{x}', \boldsymbol{\theta}^*) = \begin{pmatrix} \mathbf{V}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_1\} \\ \mathbf{V}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_2\} + \mathbf{V}_z\{(\mathbf{x}', \mathrm{D}_2)\} \end{pmatrix},$$

and

$$\mathbf{H}(\boldsymbol{\theta}^*) = \begin{pmatrix} \mathbf{H}_f(\mathrm{D}_1) & \mathbf{0} \\ \mathbf{H}_f(\mathrm{D}_2(\boldsymbol{\theta}^*)) & \mathbf{H}_z(\mathrm{D}_2) \end{pmatrix}.$$

Here $\mathbf{V}_d(\boldsymbol{\theta}^*) = \mathrm{Var}[\mathbf{d}|\boldsymbol{\theta}^*, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}, \widehat{\sigma_{\text{mes}}}, \mathbb{I}]$ is the variance of the data given in equation (2.10), where the hyperparameters have been substituted with their estimates, and $\mathbf{V}_z\{\mathbf{x}', \mathrm{D}_2\}$ has the same definition as the matrix $\mathbf{V}_{\mathbf{z}}(\mathrm{D}_2)$ as in section 2.2.2, with the inclusion of $\mathbf{x}'$ in the set of design points. The matrices $\mathbf{V}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_1\}$ and $\mathbf{V}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_2\}$ are defined the same way.

The covariance function of the true process is given by:

$$\begin{aligned} \mathrm{cov}[y(\mathbf{x}'), y(\mathbf{x}'')|\mathbf{d}, \boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\text{mes}}, \mathbb{I}] &= c_f\{(\mathbf{x}', \boldsymbol{\theta}^*), (\mathbf{x}'', \boldsymbol{\theta}^*)\} + c_z(\mathbf{x}', \mathbf{x}'') \\ &\quad - \mathbf{t}(\mathbf{x}', \boldsymbol{\theta}^*)^T\mathbf{V}_d^{-1}\mathbf{t}(\mathbf{x}'', \boldsymbol{\theta}) \\ &+ (\mathbf{h}(\mathbf{x}', \boldsymbol{\theta}^*) - \mathbf{H}(\boldsymbol{\theta}^*)^T\mathbf{V}_d^{-1}\mathbf{t}(\mathbf{x}', \boldsymbol{\theta}))^T\mathbf{W}(\boldsymbol{\theta})\{\mathbf{h}(\mathbf{x}'', \boldsymbol{\theta}) - \mathbf{H}(\boldsymbol{\theta}^*)^T\mathbf{V}_d^{-1}\mathbf{t}(\mathbf{x}'', \boldsymbol{\theta}^*)\}, \quad (2.14) \end{aligned}$$

where $\mathbf{W}(\boldsymbol{\theta}^*) = \{\mathbf{H}(\boldsymbol{\theta}^*)^T\mathbf{V}_d(\boldsymbol{\theta}^*)^{-1}\mathbf{H}(\boldsymbol{\theta}^*)\}^{-1}$.

Equations (2.13) and (2.14) are derived from the so-called predictive equations for GPs (see [Rasmussen and Williams, 2006, Chapter 2]). Since these equations are still conditioned to $\boldsymbol{\theta}^*$ which is unknown, they cannot be used directly for prediction. If we want to estimate the true process using its posterior mean for instance, we need to integrate $\mathbb{E}[y(\mathbf{x}')|\mathbf{d}, \boldsymbol{\theta}^*, \boldsymbol{\beta}, \boldsymbol{\psi}, \sigma_{\text{mes}}, \mathbb{I}]$ against the approximate posterior $\mathrm{p}(\boldsymbol{\theta}^*|\mathbf{d}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\psi}}, \widehat{\sigma_{\text{mes}}}, \mathbb{I})$ using numerical techniques. In general, the approximate posterior is not gaussian, so that the distribution of $y(\cdot)$ is not gaussian either. In the specific case where the computer model f is approximated with a linear function instead of a GP, the gaussian property is preserved.

To compute confidence intervals around the prediction of the true process, we found it best to work with a sample $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_s\}$ of the approximate posterior distribution. For each point in the sample, we can draw trajectories from the GP distribution defined by equations (2.13) and (2.14). Repeating this operation for all points $\boldsymbol{\theta}_i$ and a sufficient number of trajectories, we can compute the quantiles at each point $\mathbf{x}'$ to get the desired confidence intervals.

In uncertainty quantification studies, we sometimes have to look at the behaviour of the computer model itself and see how it compares to the data. The distribution of $f(\mathbf{x}, \boldsymbol{\theta})$ is a GP defined by the previously estimated hyperparameters $\widehat{\boldsymbol{\beta}}_f$ and $\widehat{\boldsymbol{\psi}}_f$. In the

context of the modular approach where it is assumed that the surrogate model of f is built using only $\mathbf{f}_{\text{calc}}$, the predictive equations write:

$$\mathbb{E}[f(\mathbf{x}')|\boldsymbol{\theta}^*, \boldsymbol{\beta}_f, \boldsymbol{\psi}_f, \mathbf{f}_{\text{calc}}, \mathrm{D}_1] = \mathbf{h}_f(\mathbf{x}', \boldsymbol{\theta}^*)^T \widehat{\boldsymbol{\beta}}_f + \mathbf{v}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_1\}^T \mathbf{V}_f(\mathrm{D}_1)^{-1}(\mathbf{f}_{\text{calc}} - \mathbf{H}_f \widehat{\boldsymbol{\beta}}_f), \quad (2.15)$$

and

$$\mathrm{cov}[f(\mathbf{x}'), f(\mathbf{x}'')|\boldsymbol{\theta}^*, \boldsymbol{\beta}_f, \boldsymbol{\psi}_f, \mathbf{f}_{\text{calc}}, \mathrm{D}_1] = c_f\{(\mathbf{x}', \boldsymbol{\theta}^*), (\mathbf{x}'', \boldsymbol{\theta}^*)\} - \mathbf{v}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_1\}^T \mathbf{V}_f(\mathrm{D}_1)^{-1} \mathbf{v}_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \mathrm{D}_1\}, \quad (2.16)$$

where $\mathbf{v}_f\{(\mathbf{x}, \boldsymbol{\theta}^*), \mathrm{D}_1\}$ is the column vector with i-th coefficient $c_f\{(\mathbf{x}', \boldsymbol{\theta}^*), \{\mathrm{D}_1\}_i\}$.

Once again, these equations can be integrated against the approximate posterior to obtain the predictions of the calibrated computer model. These quantities can be computed to assess the predictive capability of the computer model alone, without the correction induced by the discrepancy term.

Up until now, we have presented the general frame for KOH-style calibration and prediction. The following section will address a general point of contention in the form of the identifiability of model discrepancy and model predictions.

2.2.4 The matter of identifiability

Let us repeat the definition of the model discrepancy term according to Kennedy and O'Hagan:

$$z(\mathbf{x}) := y(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\theta}^*).$$

In this equation, there is cofounding between the model discrepancy $z(\cdot)$ and the "true parameter value" $\boldsymbol{\theta}^*$. Because the model discrepancy is arbitrary, for each possible value of parameters $\boldsymbol{\theta}$ there is a function $z(\cdot)$ that can fit the difference $y(\cdot) - f(\cdot, \boldsymbol{\theta})$, so that there is an infinite set of possible couples $(\boldsymbol{\theta}, z(\cdot))$. This is referred to as the "cofounding issue" or the "identifiability issue" and has been a major point of contention in Kennedy and O'Hagan's framework usage. This cofounding stems from the widely adopted definition of uncertainties as formulated in [Kennedy and O'Hagan, 2001a] and presented in section 2.2.1:

- **Parameter uncertainty** is the error on the model output due to the fact that the value of calibration parameters is not known exactly.
- **Model inadequacy** is the difference between the real world process and the code output at the true value of the inputs.

The definition of model error is based on a "true value" for the input parameters of the code, which is defined implicitly. This notion is later explained in the article: "The notion of a true value for $\boldsymbol{\theta}$ has the same meaning and validity as the true values of regression parameters. The true $\boldsymbol{\theta}$ is a best-fitting $\boldsymbol{\theta}$, in the sense of representing the data faithfully according to the error structure specified for the residuals.". The authors also warn that, in the case where we dispose of physical measurements for $\boldsymbol{\theta}$, these do

not necessarily match the best-fitting value. Indeed, both values hold separate meaning, and the parameter values obtained with calibration can not be directly interpreted as real, physical values.

The treatment of the identifiability issue in Kennedy and O'Hagan's original formulation is purely Bayesian [Brynjarsdóttir and O'Hagan, 2014]: all uncertainties can be learned about conjointly, by specifying an observation model for couples $(\boldsymbol{\theta}, z(\cdot))$ in the form of the likelihood function, and formulating prior distributions that are informative enough to separate both types of uncertainties. By specifying a Gaussian process model for $z(\cdot)$ with a specific covariance structure, all couples $(\boldsymbol{\theta}, z(\cdot))$ that lead to residuals abiding by this structure are regarded as plausible. However, in practical applications, it is often difficult to know if the observation models are appropriate and if the data and the prior distributions are informative enough to distinguish between model error and parameter uncertainty. Multiple authors have addressed the identifiability issue, either by keeping a Bayesian viewpoint or by trying a frequentist approach.

On the side of Bayesians, Liu [Liu et al., 2009] argue that their modular approach (see the previous section) can help in the case of high cofounding because, in the case of sequential estimation, the validity of each module can be checked. The approach of [Arendt, Apley, Chen, et al., 2012; Arendt, Apley, and Chen, 2012] is to work with vector-valued data, with multiple quantities of interest, so that the parameters and model discrepancy can be better determined. They also provide a toy example in which, when the parameter value is close to the true value, the shape of model predictions agree with the specified structure for z, and when the parameter value is far away from the reference value, the shape of model predictions change entirely. In this example, both sources of uncertainty are distinguished thanks to the Gaussian process prior. This example might seem a little contrived, but it is not uncommon that computer model predictions change shape with respect to some parameter values, especially when the model's behaviour is strongly non-linear. Another insight underlined by practitioners [Bachoc, 2013a; Carmassi et al., 2019] is that the prior mean for $z(\cdot)$ should be set to zero in hopes of reducing the cofounding with the model predictions. Besides, the prior distribution for the variance of $z(\cdot)$ is often the Jeffrey's prior that favours low values to put more weight on the model predictions that fit better the data and correspond to a low-intensity model discrepancy term. It is acknowledged that the distribution for the model discrepancy term should be not too rich so that it doesn't overshadow the computer model. Still, the distribution should be large enough to represent all the residuals observed in the analysis, and GPs provide an adequate balance to achieve this goal [Wilson and Izmailov, 2020]. Ling et al. [Ling et al., 2014] were able to propose a principled method of selection of identifiable parameters, based on a linear expansion of the computer code and the calculation of Fisher's Information Matrix. This work was extended in [Kim and Youn, 2019] to the case of hierarchically-organized models.

Tuo [Tuo and Wu, 2015a; Tuo and Wu, 2015b; Tuo and Wu, 2017] studied the frequentist properties of the KOH calibration and proved that it is not consistent in the L_2 sense, meaning that the maximum likelihood estimator of $\boldsymbol{\theta}$ does not converge in probability to the minimizer of the squared residuals. It converges instead to the minimizer of a

norm that depends on the choice of the kernel for $z(\cdot)$, showing that the effect of its prior distribution does not disappear as the number of data goes to infinity. They proposed a solution by defining the true value of parameters as the minimizer of the L_2 norm, or a weighted version of it, and they proposed a method that verifies the desired asymptotic properties. In an attempt to reconcile L_2 estimation and Bayesian calibration, [Plumlee, 2017] adapted Kennedy and O'Hagan's framework by including this external definition of the true value of parameters and employing a prior distribution for $z(\cdot)$ that is orthogonal to the gradient of the computer model, as this orthogonality condition is imposed by the external definition of the model parameters. Later [Plumlee, 2019] proposed construction of confidence intervals aimed at capturing this best parameter value. The L_2 norm is also at the center of Gu's method [Gu and Wang, 2018], whose goal is to make it the smallest possible while retaining a fully Bayesian calibration. To this end, they treated it like a variable in the fully Bayesian framework and formulated a complex prior distribution to favour the model fit to the data. In [Han et al., 2009], a separate treatment is proposed for the tuning parameters in the code (set to minimize a L_2 norm) and parameters with physical meaning, treated like in KOH's framework.

Gustafson [Gustafson, 2005] studied two alternatives when working with models suffering from cofounding: model contraction (reducing the number of parameters) and model expansion (introducing new parameters). Although not explicitly focused on the calibration of computer models, the scenarios he studied reveal that the performance of estimators can be acceptable in non-identified cases and that trying to transform the model to add more identifiability might be detrimental to the estimators' performance.

Our stance on the issue is that imposing an external definition of the best model parameter as a L_2 minimizer does not provide an acceptable solution in the case of a high model discrepancy because it amounts to finding the best-fit parameters. Even with the parameter uncertainty given by the Bayesian framework [Plumlee, 2017], the posterior distribution is necessarily centered at this value because this is the goal of the L_2-style calibration. We believe that introducing model discrepancy is instead to find model parameter values that provide predictions whose shape matches the data. On the other hand, the classical KOH framework is unsatisfying because it is often hard to estimate one shape of model discrepancy that will allow clear identifiability, using only prior information. We find that the root of the identifiability issue lies in the assumption that a "true value" exists for model parameters.

If the computer model is good enough, there might exist parameter values for which the model predictions are acceptable, meaning they are within the level of accuracy required for its use. In this case, we would have no issue calling this value "true parameter". But what about other parameter values in its neighbourhood, or even parameters that provide correct predictions but are far from the "true parameter"? Kennedy and O'Hagan's implicit assumption is that a single model discrepancy distribution is sufficient to highlight all these parameters in the posterior distribution. We argue on the contrary that, in the set of acceptable model predictions, there might be various shapes corresponding to various parameters and that a GP distribution with fixed hyperparam-

eters might not be able to recognize them all as plausible predictions. This concept is our basis for proposing another consideration of model error that results from a more flexible model discrepancy term, in section 3.1.

2.3 Calibration in practice

The general framework for KOH-style calibration has been presented, yet we have not gotten into the details of calibration that concern the first and second parts of the modular approach, which are the construction of a surrogate model and the estimation of model discrepancy. We also have not mentioned the various forms of model discrepancy terms that can be used in KOH's framework. These practical questions, addressed in this section, are at the core of applications of calibration techniques to concrete problems.

2.3.1 Construction of a surrogate model

We put the focus on the first module of the modular approach of section 2.2.3. Here we are concerned with constructing an accurate surrogate model for the computer code, a crucial concern that is often addressed in Uncertainty Quantification literature. We give in this section a review of classical techniques for surrogate construction, going from offline nDoEs to adaptive designs.

As we have previously discussed in section 2.1, the surrogate models for computer codes are primarily inspired by real experimental data analysis techniques. Notably, kriging [Matheron, 1963] that was initially applied in geostatistics. The central assumption in kriging is that the data we work with is sampled from an underlying probability distribution. Working with random fields does not suggest that the underlying phenomenon is generated from random sources. It should rather be understood as a convenient statistical framework to integrate the observed data and use it to make new predictions while providing an estimate of the uncertainty about this prediction. The quality of the regression hinges on the fact that the underlying structure of the data is correctly represented. Concerning GPs, this amounts to specifying an adequate correlation function.

The study of the best locations for experimental points is referred to as DoE for physical experiments and nDoE for computer experiments. We study two families of nDoEs: the *offline*, or *static*, and the *adaptive* designs. In the first category, the location of experiments are predetermined using a set of rules, and the experiments are run. There is a back and forth between the choice of locations and the acquisition of new observations for adaptive designs. While offline designs are more suited for DoEs, adaptive designs are relatively easy to implement for nDoEs since they require only to couple the necessary algorithms. In the following, we will assume that the input parameter space Θ is compact with dimension d.

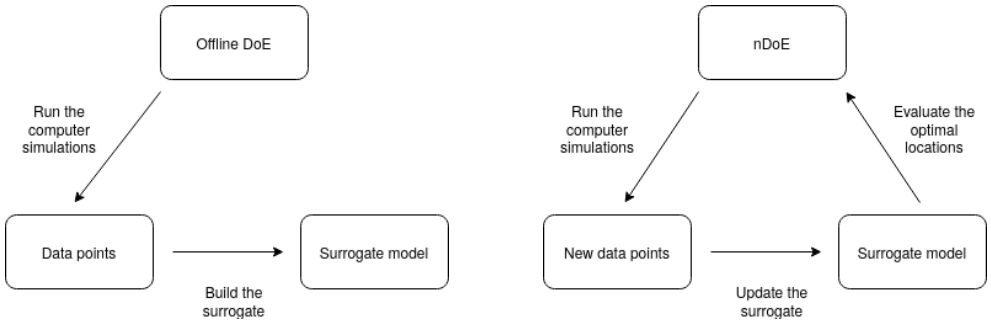


Figure 2.1: Working scheme for offline designs (left) and adaptive designs (right).

Offline designs

The first type of offline nDoEs is the *factorial designs*, where we aim to sample the whole volume of the input space while limiting the number of different values that a single input can take. This fact is justified in the case of real-world experiments, where changing the value of some inputs (temperature, flow rate, type of fluid) is time-consuming, so we would rather work with a few settings. Factorial designs can be designed to evaluate precisely the impact of individual input parameters, or couples of them. The most straightforward design is the OAT where each input parameter is set at its maximal or minimal values, with a total of 2^d points. In central composite designs, there is the addition of central points, where some input parameters are set to the center of their interval. A thorough review of factorial designs is given in [Montgomery, 2013]. Still, within the category of factorial designs, another possibility is to sample the points that correspond to integration points of quadrature rules of some polynomial families. These designs are particularly suited when the surrogate model is based on Polynomial Chaos. When the dimension of the input space is high, sparse grids can be obtained by removing points from a full factorial design [Le Maitre and Knio, 2010].

Another type of offline nDoEs is the *space-filling designs*. The objective is to populate all regions of the input space so that no effects on the response are missed. With the maximin distance criterion, the goal is to make the design points as far away as possible from one another. On the contrary, min-max designs aim to make each point in the input space as close as possible to the design points. For a detailed review of these types of designs, see [Johnson et al., 1990], and [Pronzato, 2017]. A well-known design involving random sampling is the Latin Hypercube Sampling [Mckay et al., 1979], where the input space is divided into small volumes, and the design points are randomly set into them, assuring that every row and column is filled. Another popular approach is the QMC techniques based on low-discrepancy sequences, where deterministic sequences are used to place points that resemble a sample from a uniform distribution, with better space-filling properties [Garud et al., 2017].

Some designs are constructed based on optimal criteria specifically related to the type of surrogate model used. For kriging, a popular approach is based on finding input points that maximize Shannon's entropy [Shewry and Wynn, 1987; Currin, 1988; Sebastiani and Wynn, 2000], which is named maximum entropy sampling. As the last

example, we mention Mean Squared Error-based nDoEs, where the goal is to minimize the prediction error of the surrogate model over the input space [Sacks et al., 1989].

Adaptive designs

An abundant variety of approaches aim to build a nDoE for computer experiments in an adaptive manner. In this part, we lay the grounds for such methods and review two specifically developed algorithms to answer the calibration problem.

Generally, adaptive designs are initialized with a small-size offline design on which a first surrogate model is built. Assuming we are working with a GP model, the uncertainty in the surrogate is represented by its predictive variance at a given point. Leveraging this information, it is possible to determine the regions of interest in the input space where we want to sample the computer model again so that the surrogate precision is improved where it is needed. Once the new simulations are run, the surrogate is updated, and the same technique can be applied to find new sampling points until a stopping criterion is reached. This working principle is summarized in figure 2.1. The choice of new sampling points based on the knowledge of the actual state of the surrogate model can be assimilated to *exploitation*. This strategy must be balanced with what is called *exploration* of the input space, where the goal is to sample new points far away from the observed points, based on distance or variance criterion. If there is no exploration component to the algorithm, the risk is that some regions of interest might never be discovered.

All algorithms for adaptative designs abide by the general principle that they are suited for a specific purpose determined by the context of their application. They differ in the evaluation of the quality of a surrogate, by cross-validation [Xu et al., 2014], mean squared error [Sacks et al., 1989] or mean squared error weighted with the predictive variance [Picheny et al., 2010]. The exploration strategy can also be based on the predictive variance such as Expected Improvement techniques [Jones et al., 1998; Kleijnen et al., 2012] or using the information of distance between points [Li et al., 2009]. A recent review of these algorithms can be found in [Fuhg et al., 2020], including a thorough comparison of benchmark functions. A thorny question concerns the choice of the initial offline nDoE before applying the adaptive strategy. Some insights are provided in [Leatherman et al., 2017], where the authors consider the design of both numerical and physical experiments.

Some adaptive algorithms were explicitly designed to answer the calibration problem, as is the case in [Pratola et al., 2013], where the goal of the adaptive design is to become dense in the neighbourhood of the "best parameter" defined in their article. This goal is pursued by sequentially maximizing an Expected Improvement of a likelihood factor. Later, [Damblin, Barbillon, et al., 2018] used the full posterior distribution information by defining an Expected Improvement over the KL divergence between the approximate posterior and its reference value.

2.3.2 Estimation of model discrepancy

We now address the second module of the modular approach of section 2.2.3, which is the estimation of location parameters and hyperparameters of the model discrepancy term. For ease of reading, we write equations only to estimate hyperparameters $\boldsymbol{\psi}$. The same rule treats location parameters in this approach, so if they are not null, the equations of this section remain unchanged with $\boldsymbol{\psi} \leftarrow (\boldsymbol{\beta}, \boldsymbol{\psi})$. Keeping in mind the objective of calibration, we define the objective of this step as:

Hyperparameter estimation: Find an estimator $\widehat{\boldsymbol{\psi}}$ of $\boldsymbol{\psi}$ such that $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}})$ is a good approximation to $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$.

This goal is difficult to achieve because of the marginalization equation:

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) = \int_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\psi}) \mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) \, \mathrm{d}\boldsymbol{\psi}. \tag{2.17}$$

Whereas the true distribution is obtained by averaging over the posterior distribution of location parameters and hyperparameters, our approximation is obtained by conditioning to estimated values. The first remark that we can draw from equation (2.17) is that if the posterior distribution $\mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}})$ looks like a point distribution, then the approximation is good, provided the estimate matches its mode. This fact hints at turning towards maximum likelihood type estimators rather than cross-validation techniques. We also remark that some estimators could lead to a correct approximation of $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ by chance alone, and there is not much way to know a priori whether an estimator will be successful. These considerations depend on the posterior distributions, so they are heavily application-dependent.

Let us borrow the point of view of MacKay [MacKay, 1996], for whom the act of conditioning is very unlikely in principle to produce a distribution that looks exactly like the target distribution, but that is of no consequence to the true goal. He defines the real objective of conditioning to provide a probability distribution that "puts the probability mass at the right place", meaning point values are not so important as to recognize where the plausible regions lie. The basis for his reasoning is that we are often interested in quantities that result from the integration of $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$, such as marginal distributions of parameters or integrated predictions of the model. These quantities are typically more sensitive to the location of probability "chunks" than point values.

We will now present some estimation techniques for the model discrepancy term that have appeared in the calibration literature:

- **Modular estimator from prior mean**: see [Bayarri et al., 2007; Liu et al., 2009]. The parameters' value are set to their prior mean $\widehat{\boldsymbol{\theta}}_{\mathrm{pm}}$, then the hyperparameters are estimated using the mean of the conditional posterior:

$$\widehat{\boldsymbol{\psi}}_{\mathrm{pm}} = \int_{\boldsymbol{\psi}} \boldsymbol{\psi} \, \mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{\mathrm{pm}}) \, \mathrm{d}\boldsymbol{\psi}. \tag{2.18}$$

- **Modular estimator from best fitting value**: see [Liu et al., 2009]. The parameters' value is set to the minimizer of the residuals between the model predictions and the observations $\widehat{\boldsymbol{\theta}}_{L_2}$. The model discrepancy is estimated from the residuals calculated at $\widehat{\boldsymbol{\theta}}_{L_2}$:

$$\widehat{\boldsymbol{\psi}}_{\text{bf}} = \int_{\boldsymbol{\psi}} \boldsymbol{\psi}\, \text{p}(\boldsymbol{\psi}|\mathbf{y}_{\text{obs}}, \boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{L_2})\, \text{d}\boldsymbol{\psi}. \tag{2.19}$$

- **Maximum Likelihood estimator**: obtained by maximizing the marginalized likelihood of hyperparameters:

$$\widehat{\boldsymbol{\psi}}_{\text{ml}} = \arg\max_{\boldsymbol{\psi}}\ \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\psi}}\ \int_{\boldsymbol{\theta}} \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi})\text{p}(\boldsymbol{\theta})\, \text{d}\boldsymbol{\theta}. \tag{2.20}$$

  Note that the Restricted Maximum Likelihood Estimator [Bachoc et al., 2014] works with the same principle, but with a sequential estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\psi}$.

- **Maximum A Posteriori estimator**: similar to the previous one, with the inclusion of the hyperparameter prior. This is the preferred approach in Kennedy-O'Hagan's work [Kennedy and O'Hagan, 2001b]:

$$\widehat{\boldsymbol{\psi}}_{\text{KOH}} = \arg\max_{\boldsymbol{\psi}}\ \text{p}(\boldsymbol{\psi}|\mathbf{y}_{\text{obs}}) = \arg\max_{\boldsymbol{\psi}}\ \text{p}(\boldsymbol{\psi}) \int_{\boldsymbol{\theta}} \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi})\text{p}(\boldsymbol{\theta})\, \text{d}\boldsymbol{\theta}. \tag{2.21}$$

- **LOOCV estimator**: Choose an estimator for the parameters such as $\widehat{\boldsymbol{\theta}}_{\text{pm}}$ or $\widehat{\boldsymbol{\theta}}_{L_2}$. Then, the choice of best hyperparameters is made by maximizing a score function. The classical LOOCV score function is obtained by performing calibration using all observations but one and making predictions at the hidden point to see if it matches the real data. This process is repeated for each observation point, and the LOOCV estimator is defined as:

$$\widehat{\boldsymbol{\psi}}_{\text{CV}} = \arg\max_{\boldsymbol{\psi}}\ \sum_{i=1}^{n} \log \text{p}_{\text{score}}(y_{\text{obs},i}|\mathbf{y}_{\text{obs},\sim i}, \boldsymbol{\psi}), \tag{2.22}$$

  with $y_{\text{obs},i}$ the i-th observation and $\mathbf{y}_{\text{obs},\sim i}$ all the observations expect the i-th. p_{score} is a score function that can be chosen depending if the focus is put on the prediction bias or its associated uncertainty, see [Dubrule, 1983; Bachoc, 2013b].

The cross-validation estimator was thoroughly examined in [Bachoc, 2013a], where it was proved that it is consistent and more robust than the maximum likelihood estimator in the case of misspecification of the covariance structure. This estimator is of interest for some kriging applications [Sundararajan and Keerthi, 2001; Zhang and Wang, 2010], yet to the best of our knowledge, it is only applied in calibration studies for a posteriori validation of models and not hyperparameter estimation. We consider that the estimator based on the prior mean naturally requires good prior knowledge to be effective. Using the best fit value of $\boldsymbol{\theta}$ is a reasonable approach when the model discrepancy is low but can be misleading otherwise. The Maximum Likelihood Estimator shows good asymptotic properties when the covariance structure is well-specified [Bachoc, 2013a], which we take as an incentive to try to use adequate kernels.

A Full Bayesian approach was proposed in [Higdon, Kennedy, et al., 2004], where MCMC sampling of the full posterior distribution was used so that the hyperparameters of the model discrepancy were sampled instead of being estimated. They used only three additional hyperparameters that required a MCMC of dimension 8, and they reported no mixing problems. In principle, we agree that the Full Bayesian analysis should be always performed when possible and provides more accurate results than a single point estimation. The reservations remain that it should be done within acceptable computational costs.

Some practitioners advocate that the goal of calibration is not so much to estimate posterior distributions but rather provide correct model predictions [Kejzlar et al., 2019]. The basis for this reasoning is that it is not necessary to have identified model parameters to make correct predictions [Loeppky et al., 2006]. To qualify this statement, let us remark that the parametric uncertainty estimated on one observable is transferrable to others, whereas the model discrepancy term is not. Besides, in many cases, the model and its associated uncertainties will be used for predictions outside of the range of observations, and the model discrepancy term is notoriously unsuccessful at extrapolation [Brynjarsdóttir and O'Hagan, 2014]. Accurate predictions outside of the observed domain can only rely on physical principles and sound computer models requiring well-assessed parameter posterior distributions.

2.3.3 The choice of model discrepancy term

A key point of the Bayesian calibration of Kennedy and O'Hagan lies in formulating the model discrepancy term. From the perspective of a subjective Bayesian, the mathematical properties of this term should reflect any prior knowledge available to the practitioner before seeing the data. It is thus necessary to understand the physical meaning of the prior information contained in the GP model. Furthermore, the results of calibration might be sensitive to these prior assumptions. For critics of Bayesian techniques, this is viewed as a weakness since the conclusions of the analysis heavily depend on modeling choices. The opposing viewpoint is that this is an incentive to formulate appropriate model discrepancy terms. It is also possible to employ objective priors in calibration to a certain extent. In this section, we will review a panel of model discrepancy terms that can be obtained within the KOH formulation, exploring various priors on the function itself and its hyperparameters.

From a purely subjective Bayesian point of view, the prior distribution for the model discrepancy should be formulated before seeing the observations, so before seeing how good the model is. We advocate for a more "hands-on" approach: run the computer model once or twice, see how it behaves and choose a prior adapted to what you see. Prior distributions are the tools of the Bayesian, and it would be counterproductive to refuse to use one because it is not easily translated in terms of prior information. It is also recommended to perform a sensitivity analysis by repeating the calibration with other priors. These various priors could be qualitatively compared by their marginal likelihood following the principles of Bayesian Model Selection [Robert, 2007], although computing these quantities is notoriously difficult.

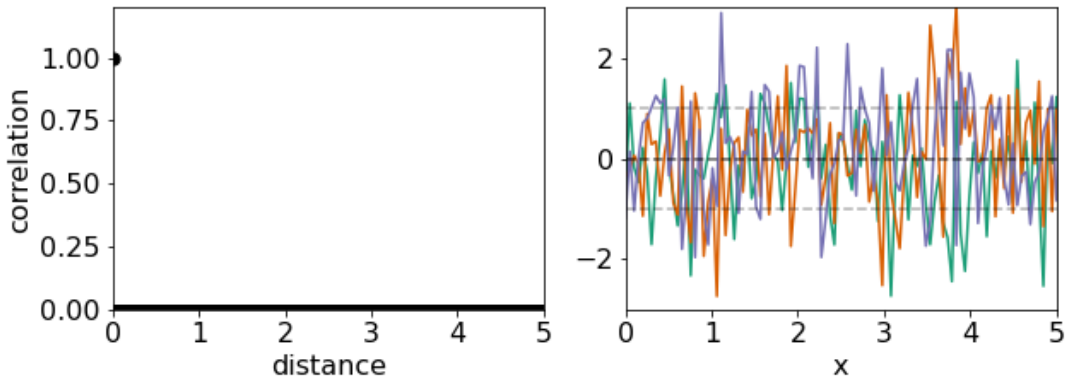


Figure 2.2: Diagonal covariance function (left) with $\sigma = 1$. Samples from the distribution (right). The dashed lines indicate the mean and one standard deviation.

In the following, we consider a GP distribution for the model discrepancy:

$$z(\cdot) \sim GP(m(\cdot), c(\cdot, \cdot)).$$

Mean function

In kriging, we usually choose among three types of structure for the mean function:

- $m(\mathbf{x}) = 0$, which means that we do not have prior beliefs that the model overpredicts or underpredicts the quantity of interest.
- $m(\mathbf{x}, \beta) = \beta$, when we believe that the model can predict the quantity of interest up to a constant β. The prior distribution on β is usually chosen uniform or gaussian.
- $m(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{h}^T(\mathbf{x})\boldsymbol{\beta}$. The general form of linear model discrepancy trend as specified in Kennedy-O'Hagan [Kennedy and O'Hagan, 2001a]. $\mathbf{h}(\cdot)$ is a vector of regression functions specified by the user. This form is adapted to the case where we have strong prior information about how the model is wrong.

In practical applications, the zero mean is often used. The reasons for this are twofold: first, specifying appropriate regression functions $\mathbf{h}(\cdot)$ is difficult and time-extensive. If we have strong prior information about the way the model is wrong, they would rather spend time improving the model. Second, there is a risk of "erasing" the model effect when using a highly structured mean function, as the equations would simplify to standard regression of the observations $\mathbf{y}_{\text{obs}}$ with the functions $\mathbf{h}(\cdot)$. This risk can be controlled by placing prior information on values of $\boldsymbol{\beta}$ that make $\mathbf{h}^T(\mathbf{x})\boldsymbol{\beta}$ small compared to $f(\mathbf{x}, \boldsymbol{\theta})$, but this is in general difficult to ensure. The zero prior mean remains a simple and effective way to perform regression, and the freedom of the statistician lies in the choice of an appropriate kernel for $z(\cdot)$.

Diagonal kernel

The diagonal kernel for the covariance function of the model discrepancy term is the following:

$$c_D(\mathbf{x}, \mathbf{x}', \sigma) = \sigma\delta(\mathbf{x}, \mathbf{x}'), \tag{2.23}$$

where $\delta(\cdot, \cdot)$ is the Dirac function. With this kernel, there is no correlation of model discrepancies for different experimental conditions $\mathbf{x}$ and $\mathbf{x}'$. In general, this correlation structure is also used for measurement error. Both errors then play a symmetric role in the likelihood function, and if they are estimated jointly, the only reliable way to distinguish them is to use dissimilar prior distributions on σ and the intensity of the measurement error σ_{mes}. For example, we suppose that the measurement error is generally lower than the model error.

Using this kernel, we also assume that the absolute values of model error are very unlikely to be higher than 2σ. The prior distribution on σ then represents our idea about the magnitude of the model error. Jeffrey's prior for this variance is the improper prior $\mathrm{p}(\sigma) \propto 1/\sigma^2$.

The constant variance conveys the assumption that the computer model's accuracy is the same in all the experimental space. In some applications, this is not necessarily the case: for example, some turbulence models are accurate in the center of the channel and not at the walls.

There is also the underlying assumption that knowing the value of model discrepancy at a given point $\mathbf{x}$ brings no information about its value at a neighbouring point $\mathbf{x}'$. Obviously, this is questionable in many applications: if the model underpredicts the quantity of interest in a given region, we expect it to give low predictions in neighbouring areas. Note that this assumption is also contained when calibrating codes without model error, which is another argument favouring a discrepancy term.

General stationary kernels

The most used kernels for model discrepancy terms are stationary, meaning that they are a function only of the lag vector between the two arguments, and we note them as such: $c(\mathbf{x}, \mathbf{x}') = c(\mathbf{x} - \mathbf{x}') = c(\mathbf{x}'')$. This is a strong assumption that we make for convenience of calculations. The resulting trajectories are often varied enough to represent the data accurately in most cases. Thus, this assumption is taken in almost all calibration studies. However, in cases where we know that the shape of model discrepancy is not the same in distinct regions of the input space, it can be worth the effort to formulate a non-stationary kernel.

Squared exponential kernel

The classical squared-exponential kernel in dimension one is defined as:

$$c_{SE}(x, \sigma, l) = \sigma^2 \exp\left(-\frac{x^2}{2l^2}\right). \tag{2.24}$$

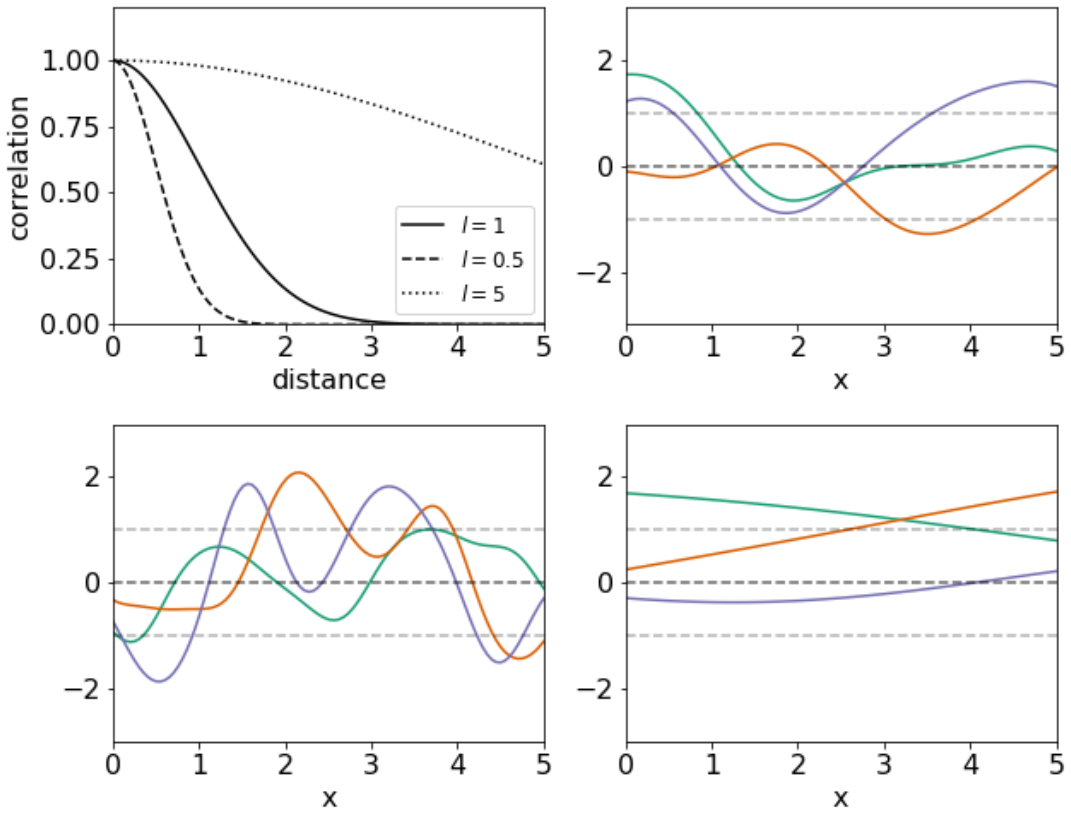


Figure 2.3: Squared exponential covariance function (top left) with $\sigma = 1$ and varying correlation lengths. Then, from left to right and top to bottom: samples with $l = 1$, $l = 0.5$ and $l = 5$. The dashed lines indicate the mean and one standard deviation.

It can be extended to dimension n as follows:

$$c_{SE,n}(\mathbf{x}, \sigma, \mathbf{l}) = \sigma^2 \exp\left(-\sum_{i=1}^{n} \frac{x_i^2}{2l_i^2}\right), \tag{2.25}$$

with $\mathbf{l}^T = (l_1, \ldots, l_n)^T$.

The hyperparameter σ has the same properties as in the diagonal kernel and represents the intensity of the model discrepancy. The hyperparameter l is interpreted as a correlation length, meaning that we believe that the knowledge of $z(x)$ brings some information about the values $z(x \pm l)$. This fact is a way to quantify the size of regions of interest in the input space.

Note that, as $l \to 0$, the squared exponential kernel becomes the diagonal kernel, and we get closer to the uncorrelated behaviour previously described. As $l \to \infty$, the correlation function is constant on the whole input space, and the trajectories of model discrepancy resemble constant functions.

Without loss of generality, assume that in the multidimensional case, the input space is $[0, 1]^n$. If the i-th correlation length l_i is significantly greater than the others, it means that the effect of the corresponding input variable is negligible with respect to the others and that the model discrepancy will not depend on its value. This property can be used in Automatic Relevance Determination (ARD) to remove irrelevant variables from the input space (see [Liu et al., 2020] for an application).

The most distinctive property of the squared exponential kernel is that the Gaussian process equipped with this kernel is almost surely infinitely differentiable, meaning that its trajectories are infinitely differentiable with probability one ([Stein, 1999]). This concept is a strong assumption to make about the model discrepancy function, which is related to the question if physical observable quantities are intrinsically continuous and differentiable, unfortunately outside of the scope of this thesis. Concerning calibration, this differentiability can be seen as a "rigidity" imposed on the functional space of model discrepancies. Intuitively, it means that if the true model discrepancy effectively belongs to the squared-exponential family (also known as the well-specified case), the rate of convergence of the model when the number of observations increases will be high. On the contrary, the squared exponential family might fail to fit the data in the misspecified case. This fact is an incentive to use kernels that provide more "flexible" trajectories, as is the case for the Matern family, which is one of the most employed kernels in GP regression.

Matern kernels

The general expression for Matern kernels in dimension one is:

$$c_{Mat}(x, \sigma, l, \nu) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{|x|}{l}\right)^{\nu} K_\nu \left(\sqrt{2\nu}\frac{|x|}{l}\right), \tag{2.26}$$

where Γ is the gamma function and K_ν is the modified Bessel function of the second kind. The new hyperparameter $\nu > 0$ controls the regularity of the trajectories obtained

in this family. As $\nu \to \infty$, the regularity is maximal, and the Matern kernel tends to the squared exponential kernel.

It is more common to work with specific values of ν, notably half-integer values. The reason being, if $\nu = p + 1/2$, the trajectories of the corresponding Matern family are almost surely continuous and p times differentiable [Rasmussen and Williams, 2006]. We then refer to Matern $1/2$ kernels, which are not differentiable, Matern $3/2$ kernels, which are differentiable once, and Matern $5/2$ kernels which are differentiable twice. Their expressions simplify greatly:

$$c_{\text{Mat}1/2}(x, \sigma, l) = \sigma^2 \exp\left(-\frac{|x|}{l}\right), \tag{2.27}$$

$$c_{\mathbf{Mat}3/2}(x, \sigma, l) = \sigma^2 \left(1 + \frac{\sqrt{3}|x|}{l}\right) \exp\left(-\frac{\sqrt{3}|x|}{l}\right), \tag{2.28}$$

$$c_{\text{Mat}5/2}(x, \sigma, l) = \sigma^2 \left(1 + \frac{\sqrt{5}|x|}{l} + \frac{5|x|^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}|x|}{l}\right). \tag{2.29}$$

Matern kernels in dimension n

There are two possible extensions of the Matern kernel in dimension n. Keeping the hypothesis that the input variables are scaled so that the imput domain is $[0, 1]^n$, the first variant is the isotropic Matern kernel,

$$c_{\text{Mat,iso}}(\mathbf{x}, \sigma, \nu, l) = c_{Mat}\left(||\mathbf{x}||, \sigma, l, \nu\right), \tag{2.30}$$

the second one is the product Matern kernel, also called tensorized or separable:

$$c_{\text{Mat,tens}}(\mathbf{x}, \sigma, \nu, \mathbf{l}) = \sigma^{2-2n} \prod_{i=1}^{n} c_{Mat}\left(x_i, \sigma, l_i, \nu\right). \tag{2.31}$$

Notable differences are that the isotropic kernel is invariant with respect to the system of coordinates. This fact means that a change in variable x_i is considered equivalent to a change in variable x_j. It is a strong assumption in general, which has the advantage of reducing the number of hyperparameters. On the contrary, tensorized kernels such as $c_{\text{Mat,tens}}$ can afford more hyperparameters because their product structure can be efficiently utilized to mitigate the curse of dimensionality [Gardner et al., 2018]. The tensored version is proposed in traditional Machine Learning libraries such as Scikit-Learn[1], GPyTorch[2] or TensorFlow[3].

[1] https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.Matern.html

[2] https://docs.gpytorch.ai/en/v1.1.1/_modules/gpytorch/kernels/matern_kernel.html

[3] https://www.tensorflow.org/probability/api_docs/python/tfp/math/psd_kernels/MaternFiveHalves

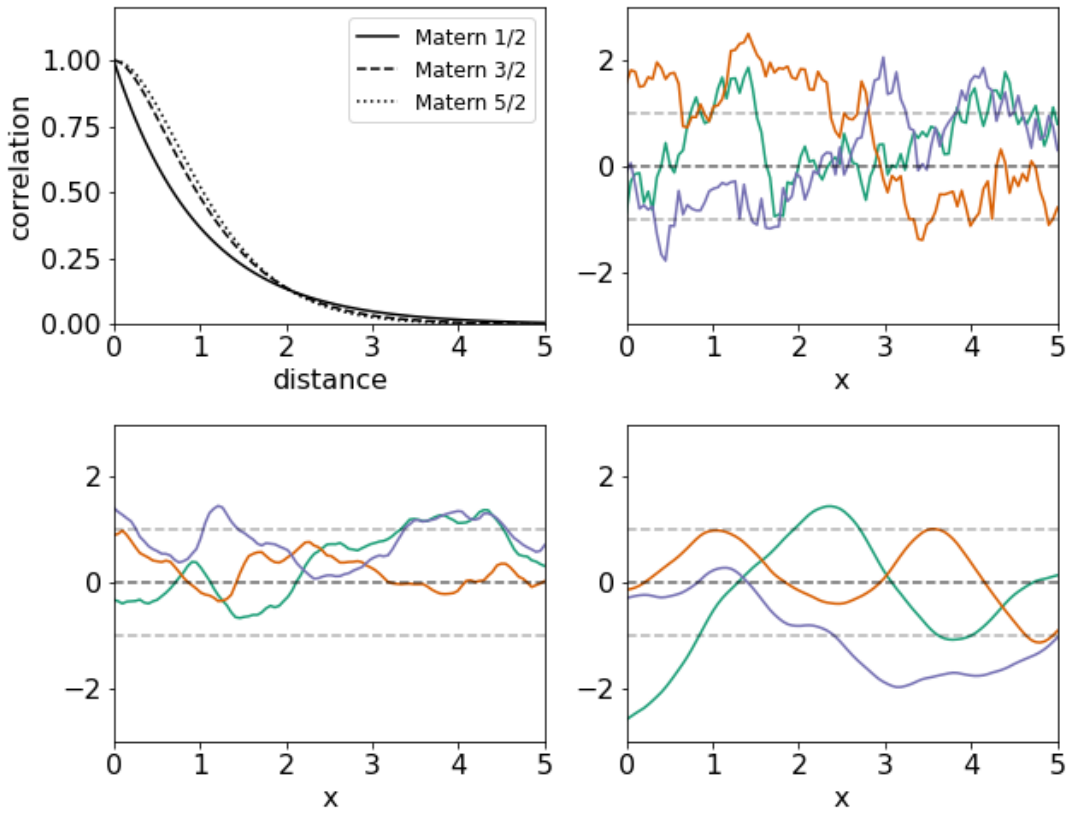


Figure 2.4: Matern covariance functions (top left) with $\sigma = 1$, $l = 1$ and varying ν. Then, from left to right and top to bottom: samples from Matern 1/2, Matern 3/2 and Matern 5/2 kernels. The dashed lines indicate the mean and one standard deviation.

Model discrepancy with specified point values

If we have prior knowledge about the value of model error in specific regions of the input space, we can quickly implement it in the Kennedy and O'Hagan framework. The same goes for derivatives of the model discrepancy. The idea is to include virtual observations in the observed dataset $\mathbf{y}_{\text{obs}}$ and perform the calibration as it is. To impose the values of derivatives, we remark that the derivatives of a GP remain a GP, so when conditioning the derivative to a specific value, the predictive distribution is still a gaussian. In the application case of [Brynjarsdóttir and O'Hagan, 2014], they formulate prior information that the model discrepancy is null at $x = 0$ and that it should be increasing in the domain (meaning they know that the model is going to be further from the truth as x increases) as a basis for this approach. This fact leads to better learning of the model parameter.

Other kernels

In this section, we have explored typical GP kernels for regression tasks with a scalar continuous output and a multidimensional input space. For further kernels belonging to this category, see [Rasmussen and Williams, 2006, Chapter 4] or the online tutorial Kernel Cookbook[4]. For an excellent theoretical insight on kernels, see [Genton, 2001]. More complex GP priors for the model discrepancy can be envisioned if the computer model falls out of this category. In the case of a strong correlation between inputs, which could occur when dealing with a lot of model parameters, it is advised to pick kernels with a product structure [Gönen and Alpaydin, 2011; Duvenaud, Lloyd, et al., 2013; Hong et al., 2017], even though, to the best of our knowledge, it has not been applied yet to calibration problems. For multi-valued outputs, i.e. to deal with plural quantities of interest, discrete or continuous, there are also dedicated works [Boyle et al., 2005; Hong et al., 2017]. Finally, in [Duvenaud, Nickisch, et al., 2011] is treated the case of additive functions, which could be used to split the high-dimensional computer model into the sum of low-dimensional functions.

2.4 Conclusion

In this section, we have presented Kennedy and O'Hagan's calibration framework to find optimal values of parameters, make corrected predictions, and perform uncertainty quantification studies on computer codes. In this framework, multiple sources of uncertainty are treated concurrently, namely parameter uncertainty, model error, measurement uncertainty and code uncertainty. This framework was largely used by practitioners and studied by statisticians, giving rise to a panel of successful applications. We have also gotten into the practical details of modelling choices, whether for the surrogate model or the model discrepancy term. The discussion on the identifiability made us question the assumption of "true value" for model parameters, which is what the model error is defined on, according to Kennedy-O'Hagan. In the next section, we will propose an alternative definition of model error that will lead us to reinterpret Kennedy and O'Hagan's

[4] https://www.cs.toronto.edu/~duvenaud/cookbook/

calibration equation. We will then propose a relevant procedure for model discrepancy estimation.

The richness of the KOH framework is to be able to accommodate a lot of real-world applications (single-valued or multi-valued, independence between outputs or strong correlation, discrete or continuous outputs, ...) and incorporate physical knowledge on the model discrepancy (positivity constraints with specified point values, specified derivatives, ...). Our point of view is that theoretical developments in GPs and calibration will be driven by the applications. Progress in this field will come from incorporating even more precise physical knowledge of the model error and dealing with more complex types of computer models. An example of a promising approach in CFD is the work of [Wu et al., 2016], who did incorporate scale factor uncertainty in the statistical formulation to extrapolate the model discrepancy term.

Chapter 3

Calibration with adaptive model discrepancy

Contents

In this chapter, we introduce the calibration framework with the adaptive model discrepancy, along with the FMP approximation that significantly reduces its numerical cost. Theoretical results include the assumptions of the suitability of the approximation, the proof of its continuity, and the proof of its robust and accurate character against the KOH calibration when the joint posterior of parameters and hyperparameters is a mixture of Gaussians with well-separated modes. Two numerical studies deal with hyperparameter estimation of the kernel in the non-microergodic case and the calibration of well-specified and misspecified models when the number of field observations tends to infinity in three asymptotic frameworks.

3.1 Calibration framework

3.1.1 The grounds for an adaptive model discrepancy

The discussions of chapter 2 helped us recognize the weaknesses in KOH's framework for model error estimation. Proposing a single-point estimation of hyperparameters (section 2.3.2) is an explicit admission that the uncertainty around the choice of prior is neglected [Kennedy and O'Hagan, 2001a]. The calibration results are sensitive to the choice of prior for the model discrepancy (section 2.3.3), and in practical applications, this prior is not obvious to formulate. Furthermore, their notion of a "true value" for the model parameters (section 2.2.4) has been called into question, and it is also the source of a lack of identifiability in the estimation (section 2.2.4). We propose an answer to these issues in the form of a new calibration framework based on an alternative consideration of model error.

Our starting point is that it is incorrect to consider that the model error can be evaluated from one model prediction that would be the best or the most accurate. Under KOH's framework, this prediction is the only one that contains information about the true process, and this is a point of contention (see the discussions in [Kennedy and O'Hagan, 2001a]). In a Bayesian framework, we argue that the whole range of predictions should be considered. Thus, we propose a new definition for the model error:

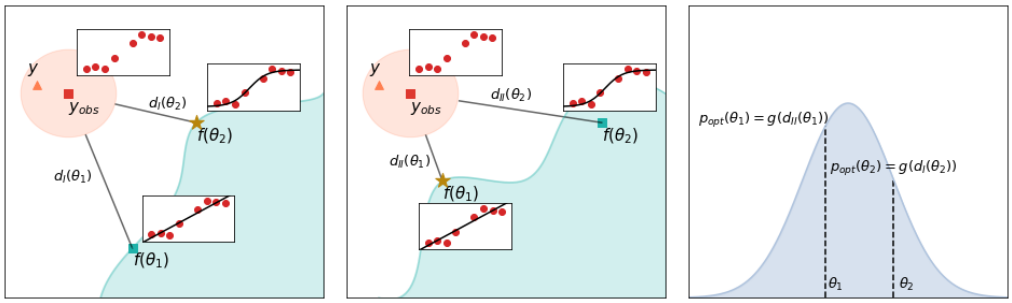
- **Model error** (or Model inadequacy) is the difference between the real world process and the posterior-averaged code output.

This definition implies that the entire calibration process plays a role in determining the model error and that no true value of input parameters exists. In KOH, the prior range of parameters is considered when estimating model discrepancy (see (2.21) and [Kennedy and O'Hagan, 2001b]), yet in their definition only the true value of parameters seems to matter. However, their estimation can lead to a distribution that is unable to faithfully represent the various shapes that the model predictions can take. We propose, instead of using a single distribution for the model discrepancy $z(\mathbf{x})$, to use an adaptive model discrepancy $z_{\boldsymbol{\theta}}(\mathbf{x})$ defined for all $(\boldsymbol{\theta}, \mathbf{x})$, as :

$$z_{\boldsymbol{\theta}}(\mathbf{x}) := y(\mathbf{x}) - f(\mathbf{x}, \boldsymbol{\theta}). \tag{3.1}$$

The idea is to compare the shapes of the model discrepancies as $\boldsymbol{\theta}$ varies and evaluate their performance through a metric that corresponds to a given GP structure. By considering multiple distributions for the model discrepancy, we can estimate the uncertainty relative to the hyperparameters of the chosen kernel. According to KOH, this is a second-order uncertainty that can be neglected in front of the parameter uncertainty. We show in our applications that, when the model predictions are sensitive to model parameters, this uncertainty cannot be neglected at the risk of giving overly confident conclusions.

The sensitivity of calibration results to the choice of model discrepancy distribution is illustrated in subfigures 3.1 (a),(b). A specific choice of distribution provides the metric under which the model predictions are evaluated. Under a metric that would consider only the fit to the data, one value of model parameters would appear to be the best (θ_2



(a) (b) (c)

Figure 3.1: Diagrams representing model error in the prediction space. y_{obs} is the vector of observations and the truth y is within the observation error (orange circle). The region of model predictions is in blue. The metric in this space is the similarity between model predictions and observations. (a): under one model error z_1, θ_2 is the true parameter of the computer model. (b): under another model error z_2, θ_1 is the true parameter. (c): our proposed correction, the posterior probability depends on the shortest distance for all envisioned model errors.

in the case of the subfigure (a)). Under some other metric, one could penalize model complexity so that θ_1 could be the preferred value (subfigure (b)). It is thus natural to ask how to choose the best metric in this problem. For Kennedy and O'Hagan, the metric is selected by considering the one that is the most likely on average over $\boldsymbol{\theta}$ (see eq. (2.21)). Our approach (subfigure (c)) is to construct a posterior distribution where the score for each $\boldsymbol{\theta}$ is calculated under its most favorable metric. Allowing a variation in metrics will allow us to capture complex variations of model predictions, keeping in mind that we must use distributions that are restrictive enough to be able to distinguish between acceptable and unacceptable model predictions.

Note that a precise definition of model error is rarely discussed in the articles in which a KOH-style calibration is performed, but it is agreed that the model discrepancy term z represents it. The KOH definition is taken for it (w.r.t. a "true value" $\boldsymbol{\theta}^*$ of model parameters); it amounts to borrowing the same definition of model error as well. In these works, the dependency of z on $\boldsymbol{\theta}^*$ is recognized, sometimes explicitly written [Bayarri et al., 2007], but does not lead to a parametric estimation of z. A variation is proposed in [Tuo and Wu, 2015a; Plumlee, 2017], in which they refine the notion of $\boldsymbol{\theta}^*$ as the minimizer of an L2 norm. By treating $\boldsymbol{\theta}^*$ as we treat $\boldsymbol{\theta}$ in this work and keeping KOH's definition, [Higdon, Kennedy, et al., 2004; Higdon, Gattiker, et al., 2008] proposed the Full Bayesian Analysis framework, which amounts to the reference calibration framework that we present below.

3.1.2 The reference calibration framework

Following eq. (3.1), the calibration equation is :

$$y_{obs}(\mathbf{x}) = f(\mathbf{x}, \boldsymbol{\theta}) + z_{\boldsymbol{\theta}}(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{3.2}$$

where notations are borrowed from section 2.2.2. For simplicity of exposition, we will consider the case where the computer model f is cheap, or a deterministic surrogate model has already been built in a previous step (see section 2.3.1). We take a GP prior for z. A linear structure is taken for the mean function :

$$z_{\boldsymbol{\theta}}(\mathbf{x})|\boldsymbol{\beta},\boldsymbol{\psi},\boldsymbol{\theta} \sim \mathrm{GP}(\mathbf{h}(\mathbf{x})^T\boldsymbol{\beta}, c_{\boldsymbol{\psi}}(\mathbf{x},\mathbf{x}')), \tag{3.3}$$

where $\mathbf{h}(\mathbf{x})(\cdot) = (h_1(\mathbf{x}),\cdots,h_p(\mathbf{x}))^T$ is a vector of p user-specified regression functions, and $c_{\boldsymbol{\psi}}$ is the covariance function. We note $\mathbf{H}$ the $n \times p$ the tendency matrix with (i,j) coefficient $h_j(\mathbf{x}_{obs,i})$, and $\mathbf{V}_{\boldsymbol{\psi}}$ the $n \times n$ covariance matrix with (i,j) coefficient $c_{\boldsymbol{\psi}}(\mathbf{x}_{obs,i},\mathbf{x}_{obs,j})$.

The observed data is normally distributed, with mean

$$\mathbb{E}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}] = \mathbf{H}\boldsymbol{\beta}, \tag{3.4}$$

and its variance is given by:

$$\mathrm{Var}[\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}] = \mathrm{Var}[\mathbf{f}_{\boldsymbol{\theta}}] + \mathbf{V}_{\boldsymbol{\psi}} + \sigma^2_{\mathrm{mes}}\mathbf{I}_n. \tag{3.5}$$

In application of Bayes' theorem, the full posterior distribution writes:

$$\mathrm{p}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\beta})\mathrm{p}(\boldsymbol{\psi})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}). \tag{3.6}$$

Now the posterior distribution for model parameters is obtained by marginalization:

$$\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) = \int_{\Psi,\mathbb{B}} \mathrm{p}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}})\ \mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}). \tag{3.7}$$

We note $\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}) := \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ to highlight its status of reference solution. We also note $\mathrm{p}_{\mathrm{KOH}}(\boldsymbol{\theta}) := \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\mathrm{KOH}}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathrm{KOH}})$ the posterior distribution obtained following the KOH approximation.

The predictive equation for the true process at a point $\mathbf{x}'$ is given by:

$$\mathrm{p}(y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}}) = \int_{\Theta,\Psi,\mathbb{B}} \mathrm{p}(y(\mathbf{x}')|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\theta},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})\ \mathrm{d}(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi}). \tag{3.8}$$

Until now, the framework is quite general and is quite similar to the KOH framework (section 2.2.2), with two differences: the model f is cheap to evaluate, and no hyperparameter estimation has been performed yet. Note that this constitutes our reference framework, and the posterior distribution (3.7) will be referred to as the "reference solution" or the "Bayesian solution" in the rest of the manuscript.

Our strategy is to compute next a parametric estimation of hyperparameters and location parameters to provide an accurate approximation to the reference solution at a reduced numerical cost.

3.1.3 Full Maximum a Posteriori estimation of model discrepancy

We introduce the dependency of $z_{\boldsymbol{\theta}}$ on $\boldsymbol{\theta}$ such that we can allow for varying distributions of $z_{\boldsymbol{\theta}}$ while keeping the GP structure with explicit predictive distributions. As mentioned in section 3.1.1, we wish to use the most favorable model discrepancy for each value of model parameters. To do so, we define optimal hyperparameters and location parameters as:

$$(\widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})) = \arg\max_{\boldsymbol{\beta},\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\beta}, \boldsymbol{\psi})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}), \tag{3.9}$$

which is the Full Maximum a Posteriori approximation, and in this thesis the subscript "FMP" will be used for any quantities related to this method, such as optimal hyperparameters or the resulting approximation of the posterior distribution. The distribution of the model discrepancy term becomes:

$$z_{\boldsymbol{\theta}}(\mathbf{x})|\boldsymbol{\beta}, \boldsymbol{\psi}, \boldsymbol{\theta} \sim \mathrm{GP}(\mathbf{h}(\mathbf{x})^T\widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), c_{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})}(\mathbf{x}, \mathbf{x}')). \tag{3.10}$$

Note that now in (3.10) the conditioning to $(\boldsymbol{\beta}, \boldsymbol{\psi})$ is redundant with the conditioning to $\boldsymbol{\theta}$, but we will keep all variables for clarity. Any classical form for the mean function or the kernel can be assumed for $z_{\boldsymbol{\theta}}$. The resulting approximation of the posterior distribution of parameters is defined as:

$$\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})) \tag{3.11}$$

We now prove that $\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta})$ is a reasonable approximation of the true posterior, based on the two approximations below:

1. For all $\boldsymbol{\theta} \in \Theta$, the distribution $\mathrm{p}(\boldsymbol{\beta}, \boldsymbol{\psi}|\boldsymbol{\theta}, \mathbf{y}_{\mathrm{obs}})$ is approximated as a point mass distribution at its mode $(\widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}))$.
2. Let $g(\boldsymbol{\theta}, \boldsymbol{\theta}')$ be defined, for $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, as :

   $$g(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}'), \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}')).$$

   We assume that, for all $\boldsymbol{\theta} \in \Theta$,

   $$\int_{\Theta} g(\boldsymbol{\theta}, \boldsymbol{\theta}')\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}')\, \mathrm{d}\boldsymbol{\theta}' \overset{\propto}{\approx} g(\boldsymbol{\theta}, \boldsymbol{\theta}). \tag{3.12}$$

The first approximation is usual in a calibration context: remark that the underlying assumption in KOH's framework is a point mass approximation of the distribution $\mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}})$ at its mode. Here we make a point mass approximation on the distributions $\mathrm{p}(\boldsymbol{\psi}|\boldsymbol{\theta}, \mathbf{y}_{\mathrm{obs}})$ so that we do not consider there exists one value of $\boldsymbol{\psi}$ that fits the error at all model predictions, rather than the errors for each model predictions can be fitted well by individual values of $\boldsymbol{\psi}$.

The second approximation is to suppose that the functions of the form $g(\boldsymbol{\theta}, \cdot)$ are similar in shape for different $\boldsymbol{\theta}$. For each function we suppose that the ratio between its posterior-averaged value and its maximal value (remark that $\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}' \in \Theta} g(\boldsymbol{\theta}, \boldsymbol{\theta}')$) does not depend on $\boldsymbol{\theta}$. This is verified for example when $\hat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})$ is a slowly varying

function of $\boldsymbol{\theta}$ (one value for optimal hyperparameters stands out), or when, for two different values of optimal hyperparameters $\hat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}_1)$ and $\hat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}_2)$, the conditional densities $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \hat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}_1))$ and $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \hat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}_2))$ are similar.

The true posterior is approximated as:

$$
\begin{aligned}
\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}) &= \int_{\Psi,\mathbb{B}} \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}})\,\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}) \\
&= \int_{\Psi,\mathbb{B}} \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}}) \left(\int_{\boldsymbol{\theta}'} \mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\theta}',\mathbf{y}_{\mathrm{obs}})\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}')\,\mathrm{d}\boldsymbol{\theta}' \right) \mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}) \\
&= \int_{\Theta} \mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}') \left(\int_{\Psi,\mathbb{B}} \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\theta}',\mathbf{y}_{\mathrm{obs}})\,\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}) \right) \mathrm{d}\boldsymbol{\theta}' \\
&\approx \int_{\Theta} \mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}')\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}'), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}'))\,\mathrm{d}\boldsymbol{\theta}' \\
&\underset{\sim}{\propto} \int_{\Theta} \mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}')\mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}'), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}'))\,\mathrm{d}\boldsymbol{\theta}' \\
&\underset{\sim}{\propto} \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})) \\
&\underset{\sim}{\propto} \mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}).
\end{aligned} \tag{3.13}
$$

The predictive density for the true process at an unobserved experimental location $\mathbf{x}'$ writes:

$$
\begin{aligned}
\mathrm{p}(y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}}) &= \int_{\Psi,\mathbb{B}} \int_{\Theta} \mathrm{p}(y(\mathbf{x}')|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\theta},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})\,\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}) \\
&= \int_{\Theta} \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) \left(\int_{\Psi,\mathbb{B}} \mathrm{p}(y(\mathbf{x}')|\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\psi},\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\boldsymbol{\beta},\boldsymbol{\psi}|\boldsymbol{\theta},\mathbf{y}_{\mathrm{obs}})\,\mathrm{d}(\boldsymbol{\beta},\boldsymbol{\psi}) \right) \mathrm{d}\boldsymbol{\theta} \\
&\approx \int_{\Theta} \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})\mathrm{p}(y(\mathbf{x}')|\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \mathbf{y}_{\mathrm{obs}})\,\mathrm{d}\boldsymbol{\theta}. \qquad (2) \\
&\approx \int_{\Theta} \mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})\mathrm{p}(y(\mathbf{x}')|\boldsymbol{\theta}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \mathbf{y}_{\mathrm{obs}})\,\mathrm{d}\boldsymbol{\theta}. \qquad (3.14)
\end{aligned}
$$

We can give a more explicit expression for the predictive density by using the predictive distributions of GPs. In the following, we note $\mathbb{E}_{\boldsymbol{\theta}}$ and $\mathrm{Var}_{\boldsymbol{\theta}}$ expectations and variances with $\boldsymbol{\theta} \sim \mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta})$. The mean of the predictive density is:

$$
\begin{aligned}
\mathbb{E}[y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}}] &= \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{E}[y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}},\boldsymbol{\theta}]] \\
&= \underbrace{\mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{x}',\boldsymbol{\theta})]}_{\text{averaged model prediction at } \mathbf{x}'} \\
&+ \underbrace{\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{t}'^{T}(\mathbf{V}_{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})} + \sigma^2_{\mathrm{mes}}\mathbf{I}_n)^{-1}(\mathbf{y}_{\mathrm{obs}} - \mathbf{f}_{\boldsymbol{\theta}} - \mathbf{H}\widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}))]}_{\text{averaged model discrepancy at } \mathbf{x}'},
\end{aligned} \tag{3.15}
$$

with $\mathbf{t}' = (c_{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})}(\mathbf{x}', \mathbf{x}_{obs,1}), \ldots, c_{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})}(\mathbf{x}', \mathbf{x}_{obs,n}))^T$. According to the formula of

total variance:

$$\begin{aligned}\mathrm{Var}[y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}}] &= \mathrm{Var}_{\boldsymbol{\theta}}[\mathbb{E}[y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}},\boldsymbol{\theta}]] + \mathbb{E}_{\boldsymbol{\theta}}[\mathrm{Var}[y(\mathbf{x}')|\mathbf{y}_{\mathrm{obs}},\boldsymbol{\theta}]] \\ &= \underbrace{\mathrm{Var}_{\boldsymbol{\theta}}[f(\mathbf{x}',\boldsymbol{\theta}) + \mathbf{t}'^{T}(\mathbf{V}_{\widehat{\psi}_{\mathrm{FMP}}(\boldsymbol{\theta})} + \sigma^2_{\mathrm{mes}}\mathbf{I}_n)^{-1}(\mathbf{y}_{\mathrm{obs}} - \mathbf{f}_{\boldsymbol{\theta}} - \mathbf{H}\widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta})))]}_{\text{uncertainty in the corrected model}} \\ &+ \underbrace{\mathbb{E}_{\boldsymbol{\theta}}[c_{\widehat{\psi}_{\mathrm{FMP}}(\boldsymbol{\theta})}(\mathbf{x}',\mathbf{x}') - \mathbf{t}'^{T}(\mathbf{V}_{\widehat{\psi}_{\mathrm{FMP}}(\boldsymbol{\theta})} + \sigma^2_{\mathrm{mes}}\mathbf{I}_n)^{-1}\mathbf{t}']}_{\text{residual uncertainty}}.\end{aligned} \tag{3.16}$$

Note that the variance formula is valid only when c is a symmetric kernel.

Looking at equation (3.16), two terms manifest our uncertainty about the real process y after calibration. The first term, named "uncertainty in the corrected model", is the variability of the model predictions after correction. It is significant when the different corrected predictions do not agree when $\boldsymbol{\theta}$ varies, which indicates that either multiple explanations of the data are plausible or the model discrepancy term is at odds with the model predictions and the data. The second term, named "residual uncertainty", is the part of the uncertainty that the corrected model cannot explain. It comes from the fact that we have only a finite number of noisy observations.

Another predictive quantity relevenant in uncertainty calibration is the calibrated computer predictions without the correction of the model discrepancy term. They can simply be obtained as:

$$\mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{x}^{'},\boldsymbol{\theta})] = \int_{\Theta} f(\mathbf{x}^{'},\boldsymbol{\theta})\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \tag{3.17}$$

for the predictive mean, and

$$\mathrm{Var}_{\boldsymbol{\theta}}[f(\mathbf{x}^{'},\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}[(f(\mathbf{x}^{'},\boldsymbol{\theta}) - \mathbb{E}_{\boldsymbol{\theta}}[f(\mathbf{x}^{'},\boldsymbol{\theta})])^2] \tag{3.18}$$

for the associated uncertainty. Both these quantities permit the evaluation of the model's predictive capacities alone, with calibrated coefficients. This aspect is crucial to check when, for example, one wishes to use the calibrated model to predict quantities that were not observed or quantities far from the observed range. Notice that, in these situations, there is no prior information about $z_{\boldsymbol{\theta}}$: thus, the only prediction we can rely on is the one of the computer model f.

In the FMP formulation, the ambiguity of the "true value" of parameters is dispelled, yet other types of identifiability problems might remain, depending on the choice of prior distributions. For example, some model predictions could be far from the observations, with a large value for the model discrepancy term, but a value for the likelihood function that is the same order as good model predictions. Thus, if the prior for $z_{\boldsymbol{\theta}}$ is not sufficiently weighted at low values, the two model predictions would have the same level of plausibility. These issues are treated by changing the prior distribution on $z_{\boldsymbol{\theta}}$. The model itself might provide the same predictions for different values of $\boldsymbol{\theta}$, which is a form of unidentifiability. A graphical representation of these types of identifiability is given in the examples of section 3.3.2.

3.1.4 Plausibility of methods for model discrepancy estimation

We now examine the optimal hyperparameters in the FMP method under the scope of Bayesian Model Comparison. For a given model M and a dataset $\mathbf{y}_{\text{obs}}$, the *marginal likelihood* $\mathrm{p}(\mathbf{y}_{\text{obs}}|M)$ is interpreted as the probability that the dataset is observed, given that the generative model is M. The marginal likelihood values are compared to determine which model is the most likely to have generated the data. We consider the prior information for hyperparameters is weak, so that the maximum a posteriori estimator simplifies to maximum likelihood: $\widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\psi})\mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi}) \approx \arg\max_{\boldsymbol{\psi}} \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi})$.

Assume that, in an alternative approximation method, the functional relationship $\hat{\boldsymbol{\psi}}_h(\boldsymbol{\theta}) = h(\boldsymbol{\theta})$ is proposed, with h a generic function. The resulting posterior for model parameters would be $\mathrm{p}_h(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi} = \boldsymbol{\psi}_h(\boldsymbol{\theta}))$. Note that in this context we also consider the single-point estimation methods such as the KOH estimator or the cross-validation estimator.

Assuming that the same parameter prior $\mathrm{p}(\boldsymbol{\theta})$ is used in all estimation methods, we have:

$$
\begin{aligned}
& \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta})) \geq \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi} = \boldsymbol{\psi}_h(\boldsymbol{\theta})), \qquad \forall \boldsymbol{\theta} \in \Theta \\
\Rightarrow \quad & \int_\Theta \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{\theta} \geq \int_\Theta \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta},\boldsymbol{\psi} = \boldsymbol{\psi}_h(\boldsymbol{\theta}))\,\mathrm{d}\boldsymbol{\theta} \\
\Leftrightarrow \quad & \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\text{FMP}}) \geq \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\psi} = \boldsymbol{\psi}_h). \qquad (3.19)
\end{aligned}
$$

Thus, the likelihood of the data when making the FMP approximation will always be higher than any other functional relationship between hyperparameters and parameters, so the FMP method will always be favored before other approximation methods from this angle. A more refined comparison could be performed by introducing priors that would penalize the approximation method complexity: indeed, if the FMP method is to be compared to the KOH approximation, for instance, in one case, the functional relationship belongs to the space of continuous functions, in the other, it lives in the space of constant functions, that is much more restricted. An appropriate prior would favor the simplicity of the KOH approximation. Unfortunately, we lack principled ways of assigning priors to methods depending on their complexity.

3.1.5 Assessing the quality of the posterior

A measure of closeness between two probability distributions is given by the KL divergence, which for two densities p and q of the random variable x is given by :

$$
D_{\text{KL}}(\mathrm{p}||\mathrm{q}) = \int_{\mathbb{X}} \mathrm{p}(x) \log \frac{\mathrm{p}(x)}{\mathrm{q}(x)}\,\mathrm{d}x. \qquad (3.20)
$$

The distribution p is often taken as the reference distribution, and q as the approximating distribution. The best approximation is reached when $\mathrm{q} = \mathrm{p}$ so that $D_{\text{KL}}(\mathrm{p}||\mathrm{q}) = 0$. Let $\tilde{\mathrm{p}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})$ be an approximation of the parameter posterior, for example obtained in the

KOH or FMP framework. Then an appropriate measure of fit is given by:

$$D_{\mathrm{KL}}(\mathrm{p}_{\mathrm{Bayes}}||\mathrm{p}_g) = \int_{\Theta} \mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) \log \frac{\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})}{\mathrm{p}_g(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})} \, \mathrm{d}\boldsymbol{\theta}. \tag{3.21}$$

Measuring the reference posterior $\mathrm{p}_{\mathrm{Bayes}}$ is required to compute this error. In most problems, we don't have access to it as it is the very goal of calibration. Due to the relatively low number of unknowns in our applications, we were able to estimate it in reasonable computer time to provide a reference solution to compare the different estimation techniques.

As a remark, note the following behaviour of the KL divergence: when the reference distribution has some probability mass in a region Θ_1 and the approximating distribution has not, the divergence takes high penalizing values. On the contrary, when the approximating distribution incorrectly guesses probability mass in a region Θ_2, the divergence is unchanged. This aspect is a kind of conservative property because it favors approximations that do not miss regions of probability mass.

3.2 Properties of the maximum a posteriori estimation

In this section, we take an interest in the problem of finding optimal hyperparameters (equation (3.9)). We discuss the existence and unicity of a solution and give the expression of the derivatives for faster optimization. Finally, we prove the continuity of the function $\boldsymbol{\theta} \mapsto \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}))$ and a regularity property on the mapping $\boldsymbol{\theta} \mapsto (\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}))$ using the Maximum Theorem.

3.2.1 The optimisation problem

We search for optimal hyperparameters and location parameters $(\boldsymbol{\psi}, \boldsymbol{\beta})$ in the space $\Psi \times \mathbb{B}$. The optimisation problem writes:

$$\begin{aligned}(\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}) &= \underset{(\boldsymbol{\psi},\boldsymbol{\beta})\in\Psi\times\mathbb{B}}{\arg\max} \; [\mathrm{L}(\boldsymbol{\psi}, \boldsymbol{\beta}) + \mathrm{p}(\boldsymbol{\psi}, \boldsymbol{\beta})] \\ &= \underset{(\boldsymbol{\psi},\boldsymbol{\beta})\in\Psi\times\mathbb{B}}{\arg\max} \left[-\frac{1}{2} \log |\mathbf{K}_{\boldsymbol{\psi}}| - \frac{1}{2} \mathbf{y}^T (\mathbf{K}_{\boldsymbol{\psi}})^{-1} \mathbf{y} + \mathrm{p}(\boldsymbol{\psi}, \boldsymbol{\beta}) \right].\end{aligned} \tag{3.22}$$

Equation (3.22) is a reparametrization of the problem 3.9, where:

- $\mathbf{y}$ is $\mathbf{y}_{\mathrm{obs}} - \mathbf{f}_{\boldsymbol{\theta}} - \mathbf{H}\boldsymbol{\beta}$,
- the likelihood criterion L is defined as $L(\boldsymbol{\psi}, \boldsymbol{\beta}) = \log(\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta})) + \frac{n}{2} * \log(2\pi)$.

Specific forms of kernel functions are considered, ensuring that $\boldsymbol{\psi} \mapsto \mathbf{K}_{\boldsymbol{\psi}}$ is continuous, and that for each $\boldsymbol{\psi} \in \Psi$, $\mathbf{K}_{\boldsymbol{\psi}}$ is symmetric positive definite. This case arrives for the Matérn family and all kernels used in this work. We also assume a continuous prior $\mathrm{p}(\boldsymbol{\psi}, \boldsymbol{\beta})$. As a consequence, the target function is continuous both in $\boldsymbol{\psi}$ and $\boldsymbol{\beta}$.

We argue that, in practice, this optimization problem is always performed in a compact search space $\Psi \times \mathbb{B}$. Lower and upper bounds can usually be formulated by looking

at the data concerning hyperparameters. Since the calibration step is done after running the computer simulations, the practitioner can plot the discrepancy between data and simulations to give reasonable upper bounds for variance hyperparameters. Hyperparameters that correspond to correlation lengths can also be limited by looking at the size of the spatial domain: the Nyquist-Shannon theorem states that signals cannot be estimated when the sampling frequency is lower than twice the signal frequency, which is an argument for setting the minimal correlation length to twice the minimal distance between observation points. Location parameters $\boldsymbol{\beta}$ can also be limited to a compact search space by considerations on the regression functions $\mathbf{h}$.

Another argument in favor of a compact search space is that, in GP regression, it is common to use priors for which the probability mass is concentrated in a compact set so that values out of this set have approximately zero posterior probability. With this argument, we consider that $\Psi \times \mathbb{B}$ is a compact space; thus, the existence of the maximum a posteriori is guaranteed following the extreme value theorem.

The unicity of the maximal value is not true under all conditions: to see this, take $\mathbf{f}_{\boldsymbol{\theta}} = \mathbf{0}$ and $\mathbf{H} = \mathbf{I}_n$. Thus at the value $\boldsymbol{\beta} = \mathbf{y}_{\text{obs}}$ the fit term $(\mathbf{y}_{\text{obs}} - \boldsymbol{\beta})^T (\mathbf{K}_{\boldsymbol{\psi}})^{-1} (\mathbf{y}_{\text{obs}} - \boldsymbol{\beta})$ is zero and the optimum in $\boldsymbol{\psi}$ is reached when the value of $|\mathbf{K}_{\boldsymbol{\psi}}|$ is lowest. The application $\boldsymbol{\psi} \mapsto |\mathbf{K}_{\boldsymbol{\psi}}|$ is not necessarily injective for classical kernel families for all datasets, thus multiple maxima might exist.

It is well-known that, in maximum likelihood estimation of hyperparameters, multiple local maxima might appear [Rasmussen and Williams, 2006]. It is also generally agreed that one maximum prevails over the others when the number of observations increases. In our applications, this behaviour was indeed observed. We thus admit in the following that, if there are enough observations, the maximum a posteriori is unique.

3.2.2 Gradients of the likelihood criterion

We suppose that the kernel function is regular with respect to its parameters, so that $\boldsymbol{\psi} \mapsto c_{\boldsymbol{\psi}}(\cdot)$ is continuous and derivable. The maximum a posteriori estimation problem then admits explicit derivatives, allowing gradient-based algorithms for faster optimization.

Following matrix-derivative formulas, we obtain the expression for the gradients with respect to hyperparameters [Rasmussen and Williams, 2006]:

$$\begin{aligned} \frac{\partial L}{\partial \psi_k}(\boldsymbol{\psi}, \boldsymbol{\beta}) &= \mathbf{y}^T \mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}_k} \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{y} - \operatorname{Tr}(\mathbf{K}_{\boldsymbol{\psi}}^{-1} \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}_k}) \\ &= \operatorname{Tr}((\boldsymbol{\alpha}\boldsymbol{\alpha}^T - \mathbf{K}_{\boldsymbol{\psi}}^{-1}) \frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}_k}), \qquad \boldsymbol{\alpha} = \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{y}. \end{aligned} \tag{3.23}$$

We have noted $\frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}_k}$ the matrix for which the (i,j) coefficient is given by: $(\frac{\partial \mathbf{K}_{\boldsymbol{\psi}}}{\partial \boldsymbol{\psi}_k})_{i,j} = \frac{\partial (\mathbf{K}_{\boldsymbol{\psi}})_{i,j}}{\partial \boldsymbol{\psi}_k}$.

The derivatives as a function of location parameters are expressed as:

$$\frac{\partial L}{\partial \beta_k}(\boldsymbol{\psi}, \boldsymbol{\beta}) = 2\mathbf{y}^T \mathbf{K}_{\boldsymbol{\psi}}^{-1} \mathbf{h}_k, \qquad \mathbf{h}_k = \begin{pmatrix} h_k(\mathbf{x}_{obs,1}) \\ h_k(\mathbf{x}_{obs,2}) \\ \dots \\ h_k(\mathbf{x}_{obs,n}) \end{pmatrix} \tag{3.24}$$

Note that the derivatives of the prior distribution must be added to the previous expression of gradients to form the gradients of the objective function in 3.22.

It is cost-effective to compute the derivatives in this optimization: the main cost lies in the evaluation of $L(\boldsymbol{\psi}, \boldsymbol{\beta})$ which requires the inversion of the matrix $\mathbf{K}_{\boldsymbol{\psi}}$. For an $n \times n$ matrix, this operation has complexity $\mathcal{O}(n^3)$ with standard algorithms such as Cholesky factorization. Once $\mathbf{K}_{\boldsymbol{\psi}}^{-1}$ is stored, the gradients can be obtained with matrix-matrix products, with complexity $\mathcal{O}(n^2)$. It is thus beneficial to include gradient information in optimization algorithms.

3.2.3 Continuity as a function of parameters

We now look at the regularity of the optimal hyperparameters with respect to the model parameters. The computer model is assumed to be continuous in its parameters so that $\boldsymbol{\theta} \mapsto \mathbf{f}_{\boldsymbol{\theta}}$ is continuous.

The main tool used in this section is the Maximum Theorem [Berge, 1963]. This theorem was proved in the case of correspondences [Aliprantis and Border, 2006], which are a generalization of functions with potentially multiple image points. We must first define these objects and give the corresponding notions of continuity.

Definition 1. *A correspondence $c : \mathbb{X} \rightrightarrows \mathbb{Y}$ is a function $f : \mathbb{X} \mapsto P(\mathbb{Y})$ where $P(\mathbb{Y})$ is the set of all subsets of $\mathbb{Y}$.*

Definition 2. *A correspondence $c : \mathbb{X} \rightrightarrows \mathbb{Y}$ is **upper-hemicontinuous** (UHC) at $x \in \mathbb{X}$ if for any open neighbourhood V of $c(x)$ there exists an open neighbourhood U of x such that for all $x' \in U$, $c(x')$ is a subset of V.*

*A correspondence $c : \mathbb{X} \rightrightarrows \mathbb{Y}$ is **lower-hemicontinuous** (LHC) at $x \in \mathbb{X}$ if for any open set V intersecting $c(x)$ there exists an open neighbourhood U of x such that $c(x^{'}) \cap V \neq \emptyset$ for each $x^{'} \in U$.*

With the following lemma we show that when the correspondence simplifies to a function, both types of continuity are equivalent to the continuity of the function:

Lemma 1. *If a correspondence $c : \mathbb{X} \rightrightarrows \mathbb{Y}$ is singleton-valued, meaning there exists a function $f : \mathbb{X} \mapsto \mathbb{Y}$ defined, for each $x \in \mathbb{X}$, as:*

$$f(x) = c(x),$$

then the three propositions are equivalent:

(i) c is upper-hemicontinuous,

(ii) c *is lower-hemicontinuous,*

(iii) f *is continuous.*

Proof. $(i) \implies (ii)$: Let $x \in \mathbb{X}$ and V an open set intersecting $c(x)$. Because c is singleton-valued, V is a neighbourhood of $c(x)$. By UHC, there exists an open neighbourhood U of x such that for all $x' \in U$, $c(x')$ is a subset of V, meaning for all $x^{'} \in U$ we have $c(x^{'}) \cap V \neq \emptyset$, thus c is LHC. $(ii) \implies (iii)$: Let $x \in \mathbb{X}$ and V an open neighbourhood of $f(x)$. By LHC, there exists an open neighbourhood U of x such that $f(U) \subset V$. Thus $f^{-1}(V) = U$ and f is continuous. $(iii) \implies (i)$: Let $x \in \mathbb{X}$ and V and open neighbourhood of $f(x)$. By continuity of f in x there exists an open neighbourhood U of x such that $U = f^{-1}(V)$, so f is UHC. □

We can now introduce the Maximum Theorem which provides a result on the regularity of correspondences arising in optimisation problems. Note that the formulation was later extended by [Tian and Zhou, 1992].

Theorem 1 (Maximum Theorem, Berge [Berge, 1963])**.** *Let* $\mathbb{X}, \mathbb{Y}$ *be topological spaces with* $\mathbb{Y}$ *compact,* $f : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$ *be a continuous real-valued function, then the correspondence* $c : \mathbb{X} \rightrightarrows \mathbb{Y}$ *defined, for each* $x \in \mathbb{X}$*, as:*

$$c(x) = \arg\max_{y \in \mathbb{Y}} f(x, y)$$

is upper-hemicontinuous and compact-valued. Besides, the function $m : \mathbb{X} \to \mathbb{R}$ *defined, for each* $x \in \mathbb{X}$*, as:*

$$m(x) = \max_{y \in \mathbb{Y}} f(x, y)$$

is continuous.

The function f in theorem 1 is our objective function $L + p$. The argument x plays the role of model parameters $\boldsymbol{\theta}$, and the argument y to be optimized over is $(\boldsymbol{\psi}, \boldsymbol{\beta})$ in our problem. This theorem states that the function $\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}), \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}}(\boldsymbol{\theta}))$ is continuous with respect to $\boldsymbol{\theta}$, no matter the shape of $\boldsymbol{\theta} \mapsto (\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}})$ as we discuss in the next paragraph.

The second result of the theorem, applied to our case, states that $\boldsymbol{\theta} \mapsto (\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}})$ is upper-hemicontinuous. In figure 3.2 we illustrate this property. In the left figure, a discontinuity occurs, and at the discontinuous point, there are two values for the optimal hyperparameter. This case might occur in applications when two explanations of the data compete, and we show an example in section 4.1.5. In the middle figure, the space of solutions becomes a closed interval, which might occur in a misspecified problem where one hyperparameter has no impact on the likelihood function. Two optimal hyperparameters remain equally likely in the right figure when the model parameter value is high. This behaviour was not observed in our applications.

If we now make the additional assumption that, for each $\boldsymbol{\theta}$, the value of $\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})$ is unique and non-empty, then $\boldsymbol{\theta} \mapsto (\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FMP}})$ is a continuous function by application of lemma 1.

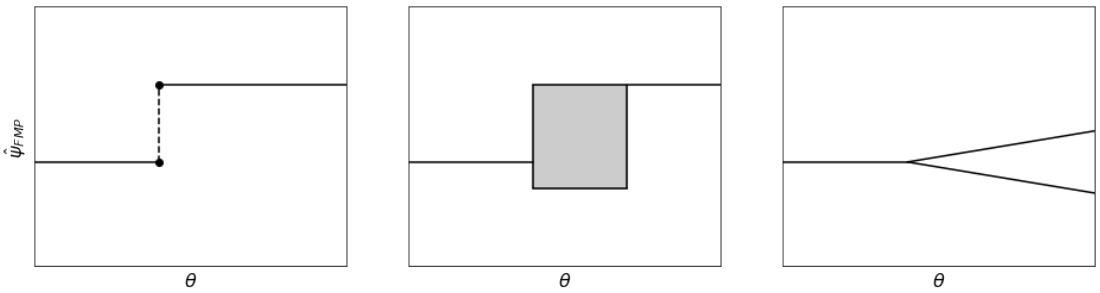


Figure 3.2: Three upper-hemicontinuous correspondences that could occur in real-world applications. Left: singleton-valued correspondence with a discontinuous change. Middle: the space of solutions becomes an interval in some region. Right: a single branch of solutions splits in two.

3.3 Asymptotic behaviour

In this section, we apply the FMP and KOH calibration techniques after performing a Laplace approximation of the joint posterior on $(\boldsymbol{\theta}, \boldsymbol{\psi})$. The results of both techniques are compared to the Bayesian solution, which is the reference. We further this study by considering a mixture of Gaussians for the joint posterior, which is assumed to represent a case where multiple explanations of the data are likely. In the two last subsections, we ask whether the joint posterior gets closer to normality when the number of observations increases. We introduce three asymptotic frameworks relevant to Kriging and give theoretical results of normality in each of them. Finally, we provide three numeric examples to study hyperparameter estimation and asymptotic frameworks.

3.3.1 Laplace approximation of the posterior

The Laplace approximation (see [MacKay, 2003] for a tutorial) is the act of transforming a hostile distribution into a pleasant gaussian. Generally, the mode of the gaussian is taken to be the mode of the actual distribution, and its variance is calculated using the second derivatives at the mode. It can also be seen as a second-order approximation of the log-probability density function around its mode.

Let us suppose that $\boldsymbol{\theta} \in \Theta$ is of dimension p and $\boldsymbol{\psi} \in \Psi$ is of dimension h. We note $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \boldsymbol{\psi})$ and suppose that $\mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}})$ is Gaussian with $\mathbb{E}[\boldsymbol{\gamma}] = \boldsymbol{\mu} = (\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{\mu}_{\boldsymbol{\psi}})$ and $\mathrm{Var}[\boldsymbol{\gamma}] = \boldsymbol{\Lambda} = \begin{pmatrix} \mathbf{V}_{\boldsymbol{\theta}} & \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T & \mathbf{V}_{\boldsymbol{\psi}} \end{pmatrix}$, which is symmetric positive definite. Following our notations, the marginal distributions are:

$$\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}} \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}}), \quad \text{and} \quad \boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}} \sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\psi}}, \mathbf{V}_{\boldsymbol{\psi}}). \tag{3.25}$$

According to the properties of the multivariate gaussian distribution, the density of $\boldsymbol{\theta}$

conditioned to a value of $\boldsymbol{\psi}$ is given by:

$$\begin{aligned}\boldsymbol{\theta}|\boldsymbol{\psi}, \mathbf{y}_{\text{obs}} &\sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{\psi}}, \mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\theta}}),\\ \boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{\psi}} &= \boldsymbol{\mu}_{\boldsymbol{\theta}} + \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}}^{-1}(\boldsymbol{\psi} - \boldsymbol{\mu}_{\boldsymbol{\psi}}),\\ \mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}} &= \mathbf{V}_{\boldsymbol{\theta}} - \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}.\end{aligned} \tag{3.26}$$

A specificity of Gaussian distributions is that $\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}$ does not depend on the value of $\boldsymbol{\psi}$ it is conditioned to. The proof of equation (3.26) is obtained by the definition of conditional probability and "completing the square" in the exponential function. The expression of the conditional distribution of hyperparameters is, by symmetry:

$$\begin{aligned}\boldsymbol{\psi}|\boldsymbol{\theta}, \mathbf{y}_{\text{obs}} &\sim \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\psi}|\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}}),\\ \boldsymbol{\mu}_{\boldsymbol{\psi}|\boldsymbol{\theta}} &= \boldsymbol{\mu}_{\boldsymbol{\psi}} + \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}),\\ \mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}} &= \mathbf{V}_{\boldsymbol{\psi}} - \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T.\end{aligned} \tag{3.27}$$

One remark is to be made about the *reduction of variance* property in the conditional distribution. In the simple case $p = h = 1$, we have $v_{\boldsymbol{\theta}|\boldsymbol{\psi}} = v_{\boldsymbol{\theta}} - \frac{c_{\boldsymbol{\theta},\boldsymbol{\psi}}^2}{v_{\boldsymbol{\psi}}} \leq v_{\boldsymbol{\theta}}$, where lowercase quantities represent the scalar versions of the variance-covariance matrices. This shows that the conditional is narrower than the marginal. In general dimension, under the Loewner order for matrices we have $\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}} \leq \mathbf{V}_{\boldsymbol{\theta}}$ because $\mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}$ is symmetric positive definite.

The likelihood function writes:

$$\mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\text{obs}}) = \frac{1}{(2\pi)^{(p+h)/2}\sqrt{|\boldsymbol{\Lambda}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}^{-1}(\boldsymbol{\gamma} - \boldsymbol{\mu})\right). \tag{3.28}$$

We now calculate the posterior distribution of parameters following the three calibration techniques.

The exact marginal distribution of the parameters is:

$$\mathrm{p}_{\text{Bayes}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}}). \tag{3.29}$$

KOH approximation

The first step is to compute the marginal of hyperparameters and get its maximum. The maximum of the marginal is immediately given by:

$$\hat{\boldsymbol{\psi}}_{\text{KOH}} = \boldsymbol{\mu}_{\boldsymbol{\psi}}. \tag{3.30}$$

And following the KOH framework:

$$\begin{aligned}\mathrm{p}_{\text{KOH}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) &= \mathrm{p}(\boldsymbol{\theta}|\hat{\boldsymbol{\psi}}_{\text{KOH}}, \mathbf{y}_{\text{obs}})\\ &= \mathrm{p}(\boldsymbol{\theta}|\boldsymbol{\mu}_{\boldsymbol{\psi}}, \mathbf{y}_{\text{obs}})\\ &= \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}|\boldsymbol{\mu}_{\boldsymbol{\psi}}}, \mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}})\\ &= \mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}).\end{aligned} \tag{3.31}$$

Note the *reduction of variance* property: because the matrix $\mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}$ is symmetric positive definite, under the Loewner order we have $\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}} \leq \mathbf{V}_{\boldsymbol{\theta}}$. This order can be understood as the following: let $\{\lambda_i\}_{1\leq i\leq p}$ be the eigenvalues of $\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}$ in a decreasing order and $\{\lambda_i^{'}\}_{1\leq i\leq p}$ be the eigenvalues of $\mathbf{V}_{\boldsymbol{\theta}}$ in a decreasing order. Then, for all $i \leq p$, we have $\lambda_i \leq \lambda_i^{'}$.

It is seen that the approximated posterior of $\boldsymbol{\theta}$ remains Gaussian, with correct mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$, but a reduced covariance compared to the Bayes solution.

FMP approximation

To proceed with the FMP estimation, the first step is to estimate the optimised hyper-parameters by solving the optimisation problem :

$$\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\psi})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta},\boldsymbol{\psi}) = \arg\max_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\psi}|\boldsymbol{\theta},\mathbf{y}_{\mathrm{obs}}).$$

And we get directly from the expression of the conditional density 3.27:

$$\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) = \boldsymbol{\mu}_{\boldsymbol{\psi}|\boldsymbol{\theta}} = \boldsymbol{\mu}_{\boldsymbol{\psi}} + \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}). \tag{3.32}$$

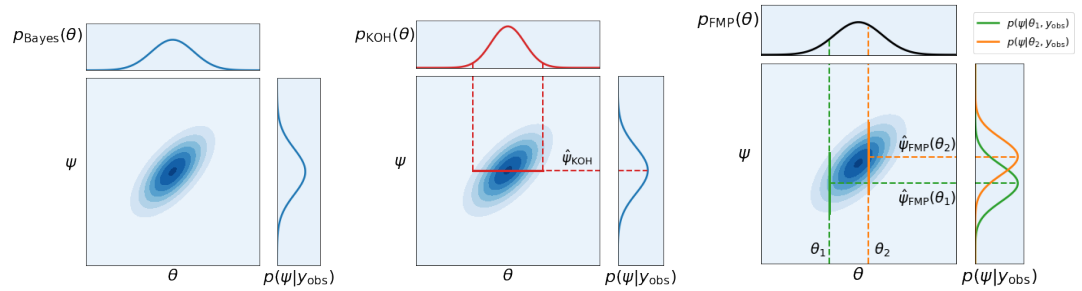
Thus the optimal hyperparameters depend linearly on the parameters. To evaluate the approximate posterior distribution of parameters, we need to express the inverse of the covariance matrix Λ, using the Matrix Block Inversion Lemma [Rasmussen and Williams, 2006]:

$$\boldsymbol{\Lambda}^{-1} = \begin{pmatrix} \mathbf{V}_{\boldsymbol{\theta}}^{-1} + \mathbf{V}_{\boldsymbol{\theta}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} & -\mathbf{V}_{\boldsymbol{\theta}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}}^{-1} \\ -\mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}}^{-1} \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} & \mathbf{V}_{\boldsymbol{\psi}|\boldsymbol{\theta}}^{-1} \end{pmatrix}, \tag{3.33}$$

Then, we have:

$$\begin{aligned} \mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) &\propto \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})|\mathbf{y}_{\mathrm{obs}}) \\ &\propto \exp\left(-\frac{1}{2} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \\ \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{\boldsymbol{\psi}} \end{pmatrix}^T \boldsymbol{\Lambda}^{-1} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \\ \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) - \boldsymbol{\mu}_{\boldsymbol{\psi}} \end{pmatrix} \right) \\ &\propto \exp\left(-\frac{1}{2} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \end{pmatrix}^T \boldsymbol{\Lambda}^{-1} \begin{pmatrix} \boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}} \\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \end{pmatrix} \right) \\ &\propto \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \begin{pmatrix} \mathbf{1}_{p\times p} \\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} \end{pmatrix}^T \boldsymbol{\Lambda}^{-1} \begin{pmatrix} \mathbf{1}_{p\times p} \\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{\boldsymbol{\theta}}^{-1} \end{pmatrix} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \right) \\ &\propto \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}})^T \mathbf{V}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta}}) \right). \end{aligned} \tag{3.34}$$

The density is proportional to the exponential of a quadratic form, so it is Gaussian, and its mean and covariance matrix corresponds to the true marginal distribution. In this case, the FMP approximation is exact.



(a) Bayesian solution (b) KOH estimation (c) FMP estimation

Figure 3.3: Joint posterior of parameters and hyperparameters as a single-mode gaussian. All probabilities are posterior probabilities, implicitely conditioned to the observations $\mathbf{y}_{\text{obs}}$.

Discussion

The behaviour of all three methods is shown on the illustrative figure 3.3, with $\dim(\boldsymbol{\theta}) = \dim(\boldsymbol{\psi}) = 1$. In the reference solution 3.3a, both marginals for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are plotted. The KOH estimation 3.3b is done by computing the true marginal of $\boldsymbol{\psi}$ and making a projection of the joint posterior on the plane that corresponds to $\hat{\boldsymbol{\psi}}_{\text{KOH}}$. In the FMP estimation 3.3c, the marginals are computed at each value of $\boldsymbol{\theta}$ and the parameter posterior probability is proportional to the value of the joint probability at the point $(\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}))$.

Measure of fit

The quality of the KOH approximation is given as the KL divergence between two multivariate gaussians with identical mean:

$$\begin{aligned} D_{\text{KL}}(\text{p}_{\text{Bayes}}||\text{p}_{\text{KOH}}) &= D_{\text{KL}}(N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}})||N(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}})) \\ &= \frac{1}{2}\left(\log\frac{|\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}|}{|\mathbf{V}_{\boldsymbol{\theta}}|} + \text{Tr}(\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}^{-1}\mathbf{V}_{\boldsymbol{\theta}}) - p\right) \end{aligned} \tag{3.35}$$

In the case $p = h = 1$, we get the expression:

$$D_{\text{KL}}(\text{p}_{\text{Bayes}}||\text{p}_{\text{KOH}}) = \frac{1}{2}\left(\log(1-d) + \frac{d}{1-d}\right), \tag{3.36}$$

with $d = \frac{(c_{\boldsymbol{\theta},\boldsymbol{\psi}})^2}{v_{\boldsymbol{\theta}} v_{\boldsymbol{\psi}}}$. Note that we have $0 \leq d \leq 1$ because $\boldsymbol{\Lambda}$ is positive definite. The quantity d is a normalized measure of the posterior covariance between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. On figure 3.4, it can be seen that the "false certitude" effect is stronger in the case of high posterior correlation.

3.3.2 Approximation as a mixture of gaussians

We now consider the case where the joint posterior $\text{p}(\boldsymbol{\gamma}|\mathbf{y}_{\text{obs}})$ is a mixture of Gaussians with well-separated modes. This assumption represents the cases where the space of

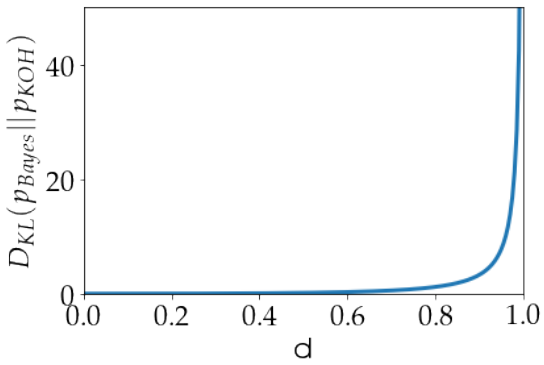


Figure 3.4: Precision of the KOH method as a function of the intensity of the posterior correlation between parameters and hyperparameters in the 1D case.

model parameters is divided into regions, and the optimal model error distribution is different for each region.

Assume that the joint density is a mixture of m Gaussians, with weights $(\pi_i)_{i\leq m}$, such that $\sum_{i=1}^{m}\pi_i = 1$. Their means are $\boldsymbol{\mu}_i = \begin{pmatrix}\mu_{\boldsymbol{\theta},i}\\ \mu_{\boldsymbol{\psi},i}\end{pmatrix}$ and their covariance matrices: $\boldsymbol{\Lambda}_i = \begin{pmatrix}\mathbf{V}_{\boldsymbol{\theta},i} & \mathbf{C}^T_{\boldsymbol{\theta},\boldsymbol{\psi},i}\\ \mathbf{C}_{\boldsymbol{\theta},\boldsymbol{\psi},i} & \mathbf{V}_{\boldsymbol{\psi},i}\end{pmatrix}$. The joint density writes:

$$\begin{aligned}\mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}) &= \sum_{i=1}^{m}\pi_i \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}},\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)\\ &= \frac{1}{(2\pi)^{(p+h)/2}}\sum_i \frac{\pi_i}{\sqrt{|\boldsymbol{\Lambda}_i|}}\exp\left(-\frac{1}{2}(\boldsymbol{\gamma}-\boldsymbol{\mu}_i)^T\boldsymbol{\Lambda}_i^{-1}(\boldsymbol{\gamma}-\boldsymbol{\mu}_i)\right).\end{aligned} \tag{3.37}$$

We make the hypothesis that the m Gaussians are well separated on the spaces Θ and Ψ. Geometrically, the well-separated hypothesis on Θ means that when projecting the Gaussians on the first p coordinates, the 95% confidence ellipses of each gaussian do not intersect each other. We can write this as:

"The intervals $\Theta_j = \{\boldsymbol{\theta} \text{ s. t. } (\boldsymbol{\theta}-\boldsymbol{\mu}_j)^T\mathbf{V}^{-1}_{\boldsymbol{\theta},j}(\boldsymbol{\theta}-\boldsymbol{\mu}_j) \leq t_{95}(p)\}$ are disjoint for $j \leq m$,"

where we have noted $t_{95}(p)$ the 95% quantile of the χ_2 law with p degrees of freedom. We suppose the equivalent condition for the projections over Ψ.

We also make the assumption that the weights $\{\pi_i\}_{1\leq i\leq m}$ are each of same order of magnitude, so that, for $1 \leq i \leq m$, if $(\boldsymbol{\theta},\boldsymbol{\psi})$ is close to $\boldsymbol{\mu}_i$ then $\mathrm{p}(\theta,\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) \approx \pi_i \mathrm{p}(\theta,\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}},\boldsymbol{\mu}_i,\boldsymbol{\Lambda}_i)$.

By linearity, the true marginal density of the parameters is the linear combination

of marginals:

$$\begin{aligned} \mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}) &= \int_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}})\,\mathrm{d}\boldsymbol{\psi} = \sum_{i=1}^{m} \pi_i \int_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)\,\mathrm{d}\boldsymbol{\psi} \\ &= \frac{1}{(2\pi)^{p/2}} \sum_{i=1}^{m} \frac{\pi_i}{\sqrt{|\mathbf{V}_{\boldsymbol{\theta},i}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{\boldsymbol{\theta},i})^T \mathbf{V}_{\boldsymbol{\theta},i}^{-1}(\boldsymbol{\theta} - \mu_{\boldsymbol{\theta},i})\right). \end{aligned} \tag{3.38}$$

KOH approximation

First, the hyperparameter marginal is computed:

$$\mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) = \frac{1}{(2\pi)^{h/2}} \sum_{i=1}^{m} \frac{\pi_i}{\sqrt{|\mathbf{V}_{\boldsymbol{\psi},i}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\psi} - \boldsymbol{\mu}_{\boldsymbol{\psi},i})^T \mathbf{V}_{\boldsymbol{\psi},i}^{-1}(\boldsymbol{\psi} - \mu_{\boldsymbol{\psi},i})\right). \tag{3.39}$$

According to our separation hypothesis over Ψ, the maximum of $\mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}})$ is:

$$\hat{\boldsymbol{\psi}}_{\mathrm{KOH}} = \arg\max_{\boldsymbol{\psi}} \mathrm{p}(\boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) = \boldsymbol{\mu}_{i_{KOH}} \quad \text{with } i_{\mathrm{KOH}} \text{ s.t. } i_{\mathrm{KOH}} = \arg\max_{i \leq m} \frac{\pi_j}{\sqrt{\mathbf{V}_{\boldsymbol{\psi},i}}}. \tag{3.40}$$

And the KOH solution is a single-mode multivariate Gaussian:

$$\begin{aligned} \mathrm{p}_{\mathrm{KOH}}(\boldsymbol{\theta}) &= \mathrm{p}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\psi} = \boldsymbol{\mu}_{i_{\mathrm{KOH}}}) \\ &= \mathrm{N}(\boldsymbol{\mu}_{i_{\mathrm{KOH}},\boldsymbol{\theta}}, \mathbf{V}_{i_{\mathrm{KOH}},\boldsymbol{\theta}|\boldsymbol{\psi}}) \end{aligned} \tag{3.41}$$

with:

$$\mathbf{V}_{i_{\mathrm{KOH}},\boldsymbol{\theta}|\boldsymbol{\psi}} = \mathbf{V}_{i_{\mathrm{KOH}},\boldsymbol{\theta}} - \mathbf{C}_{i_{\mathrm{KOH}},\boldsymbol{\theta},\boldsymbol{\psi}}^T \mathbf{V}_{i_{\mathrm{KOH}},\boldsymbol{\psi}} \mathbf{C}_{i_{\mathrm{KOH}},\boldsymbol{\theta},\boldsymbol{\psi}}.$$

The solution of the KOH estimation is a single-mode Gaussian with a reduced variance matrix. Besides, the selection of the mode is driven by the criteria exhibited on i_{KOH}, so it does not necessarily correspond to the true maximum of the Bayesian solution.

FMP approximation

According to the separation hypothesis, the confidence intervals $\{\Theta_i\}_{i \leq m}$ are disjoint so that:

$$\begin{aligned} \text{for} \quad \boldsymbol{\theta} \in \Theta_i, \quad & \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}) \approx \pi_i \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i), \\ \text{and} \quad & \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) = \hat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta}) := \boldsymbol{\mu}_{i,\boldsymbol{\psi}} + \mathbf{C}_{i,\boldsymbol{\theta},\boldsymbol{\psi}} \mathbf{V}_{i,\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}}). \end{aligned} \tag{3.42}$$

We note $\mathrm{p}^*_{\mathrm{FMP}}(\boldsymbol{\theta}) = \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\psi} = \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})|\mathbf{y}_{\mathrm{obs}}) = \sum_{j=1}^{m} \pi_j \mathrm{p}(\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j)$ the unnormalized FMP approximation of the parameter posterior. We have the following two properties:

1. Dominance of the i-th mode in the interval Θ_i:

   $$\forall \boldsymbol{\theta} \in \Theta_i, \quad \pi_i \mathrm{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta})|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) \gg \sum_{j \neq i} \pi_j \mathrm{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta})|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j).$$

   This is the application of the separation property in Ψ and the low discrepancy in the weights $\{\pi_i\}_{1 \leq i \leq m}$.

2. Optimal hyperparameter in Θ_i:

$$\pi_i \mathrm{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta}) | \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i) = \frac{\pi_i \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})^T \mathbf{V}_{i,\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})\right)}{(2\pi)^{(p+h)/2}\sqrt{|\boldsymbol{\Lambda}_i|}}.$$

This was obtained in the unimodal case (equation (3.34)).

Thus, a natural approximation for $\mathrm{p}^*_{\mathrm{FMP}}$, for $\boldsymbol{\theta} \in \Theta$, is:

$$\mathrm{p}^*_{\mathrm{FMP}}(\boldsymbol{\theta}) \approx \sum_{j=1}^{m} \pi_j \mathrm{p}(\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}_{j,\mathrm{FMP}}(\boldsymbol{\theta}) | \mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_j, \boldsymbol{\Lambda}_j), \tag{3.43}$$

and if $\boldsymbol{\theta} \in \Theta_i$ it simplifies to:

$$\mathrm{p}^*_{\mathrm{FMP}}(\boldsymbol{\theta}) \approx \frac{\pi_i}{(2\pi)^{(p+h)/2}\sqrt{|\boldsymbol{\Lambda}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})^T \mathbf{V}_{i,\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})\right). \tag{3.44}$$

Since most of the probability mass of $\mathrm{p}^*_{\mathrm{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ is contained within the intervals Θ_i, we have:

$$\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}) = \frac{1}{K}\sum_{i=1}^{m} \frac{\pi_i}{\sqrt{|\boldsymbol{\Lambda}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})^T \mathbf{V}_{\boldsymbol{\theta},i}^{-1}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})\right) \quad \text{for } \boldsymbol{\theta} \in \Theta, \tag{3.45}$$

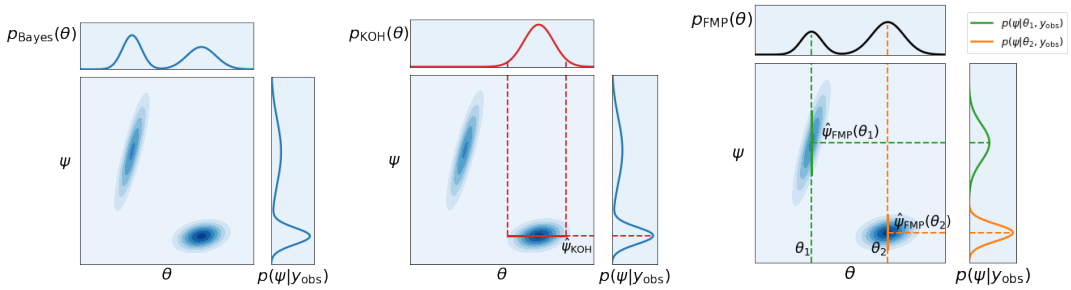
with the normalizing constant:

$$\begin{aligned} K &= \int_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{m} \frac{\pi_i}{\sqrt{|\boldsymbol{\Lambda}_i|}} \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\mu}_{i,\boldsymbol{\theta}})^T \mathbf{V}_{\boldsymbol{\theta},i}^{-1}(\boldsymbol{\theta} - \mu_{i,\boldsymbol{\theta}})\right) \mathrm{d}\boldsymbol{\theta} \\ &\approx (2\pi)^{(p/2)} \sum_{i=1}^{m} \pi_i \frac{\sqrt{|\mathbf{V}_{\boldsymbol{\theta},i}|}}{\sqrt{|\boldsymbol{\Lambda}_i|}}. \end{aligned}$$

The distribution $\mathrm{p}_{\mathrm{FMP}}$ is a linear combination of the m marginals that correspond to the original Gaussian distributions. The variances of the different peaks are correctly estimated, but there is bias in the estimation of the relatives weights of the peaks: where in the reference solution the weights are the $\{\pi_i\}_{i\leq m}$, here the weights are the $\{\pi'_i\}_{i\leq m} = \{\frac{1}{k}\pi_i \frac{\sqrt{|\mathbf{V}_{\boldsymbol{\theta},i}|}}{\sqrt{|\boldsymbol{\Lambda}_i|}}\}_{i\leq m}$ (with k the normalizing constant such that $\sum_{i=1}^m \pi'_i = 1$). To give some physical sense about this distortion of the weights, reduce to dimension one and take the simple example where $c_{i,\boldsymbol{\theta},\boldsymbol{\psi}} = 0$ for all i (no posterior correlation for all modes). Then the family of weights estimated by the FMP method becomes $\{\frac{\pi_i}{\sqrt{v_{i,\boldsymbol{\psi}}}}\}_{i\leq m}$ (up to a normalizing constant), which shows that the modes with high hyperparameter variance are penalized.

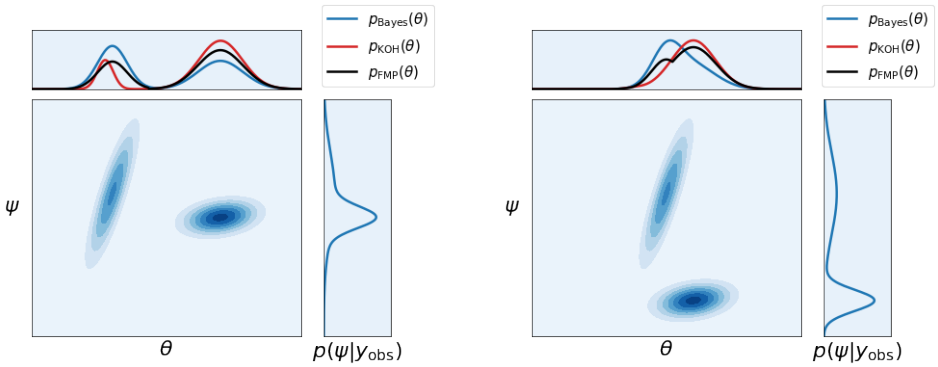
Without clear separation of the modes

An illustration of all three methods is given in figure 3.5. In this example, the KOH estimation finds only one peak that does not necessarily correspond to the true maximum of the posterior, so the conclusions about $\boldsymbol{\theta}$ are potentially misleading. In the FMP



(a) Bayesian solution (b) KOH estimation (c) FMP estimation

Figure 3.5: Joint posterior of parameters and hyperparameters as a mixture of gaussians.



(a) Intersection of projections in Ψ. (b) Intersection of projections in Θ.

Figure 3.6: Gaussian mixture without clear separation of the modes.

estimation, all the peaks are found, yet their relative weights might be wrongly estimated, which can lead to an inversion of the importance of the peaks so that the maximum a posteriori estimator is incorrect, as shown on the subfigure 3.5c. This characteristic will be found in any method that uses estimated hyperparameters instead of marginalization because volume effects cannot be seen. We argue that it is still favorable to perform the FMP estimation in this case so that no possible explanation of the data is missed.

We now discuss what happens when the well-separated hypothesis is not respected. On figure 3.6a, the projections of the modes on the space Ψ intersect each other. This case occurs when a single model discrepancy adequately represents multiple regions of the model parameters. This case is favorable because both KOH and FMP approximations will see the posterior modes.

If we now take the case where the projections of the modes are not well-separated in Θ, as is shown in figure 3.6b, the situation is unfavourable for both estimation methods since the hyperparameter marginals can not be reliably estimated as point masses. As a consequence, some modes will not appear on the parameter posterior. The parameter region will still be identified as plausible, but some of the probability mass will be overlooked. This situation happens when the distribution of the model discrepancy is at odds with the data, and no values of hyperparameters stand out. It points to the need for another distribution of $z_{\boldsymbol{\theta}}$.

Measure of fit

We briefly go back to the well-separated case. The fit of the KOH approximation measured by the KL divergence is almost infinity because some modes are missed (see remark in section 3.1.5). It is simple to get an expression for the fit of the FMP method:

$$\begin{aligned} D_{\mathrm{KL}}(\mathrm{p}_{\mathrm{Bayes}}||\mathrm{p}_{\mathrm{FMP}}) &= D_{\mathrm{KL}}(\sum_{i=1}^{m} \pi_i \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)|| \sum_{i=1}^{m} \pi'_i \mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\mathrm{obs}}, \boldsymbol{\mu}_i, \boldsymbol{\Lambda}_i)) \\ &\approx \sum_{i=1} \pi_i \log \frac{\pi_i}{\pi'_i}. \end{aligned} \tag{3.46}$$

We have used once again the approximation that most of the probability mass is contained in the intervals Θ_i. It results that the measure of fit amounts to the discrete KL divergence between the two distributions of weights.

3.3.3 Theoretical results in asymptotic regimes

We have seen that the normality of the posterior is a favorable condition for our estimation method. We now ask whether it is common in calibration problems to find this normality, and intuitively we consider cases where the number of observations is high.

There are two separate approaches to asymptotics, be it the frequentist way or the Bayesian way, with some bridges between the two. Frequentists study properties of *estimators*, for example, consistency or asymptotic normality. These concepts in the

Bayesian field are related to consistency or asymptotic normality of the *posterior distribution*: does it shrink around a specific value, or does it look like a gaussian distribution? Theorems related to the Bayesian properties are often called Bernstein-Von Mises results.

We present the two classical asymptotic regimes for Kriging: increasing domain and fixed domain. Since they deal with acquiring new observations that are *not independent* from the previous ones, Bernstein-Von Mises results can not be applied directly. Thus, we will present frequentist results for those regimes. We also introduce a new asymptotic regime that we name repeated domain, in which we work with independent trajectories so that Bernstein-Von Mises results are applicable. The discussion will be completed with numerical examples in section 3.3.4, to link theoretical asymptotic properties and practical estimation.

The asymptotic regimes are introduced as follows:

- **ID:** this regime corresponds to the acquisition of new observations in an expanding domain of $\mathbf{x}$. Generally, it is applied to Kriging when $\mathbf{x}$ represents a time variable, so the observations stem from a time series that is regularly sampled.
- **FD:** this regime, also referred to *infill asymptotics* in literature, corresponds to acquiring new observations inside the same domain of $\mathbf{x}$, so that the observation points tend to become dense in the domain. This is the typical asymptotic regime when $\mathbf{x}$ is a space variable, and the design of experiments is gradually refined to learn the variations of the function inside the domain.
- **RD:** this regime corresponds to experiments that have a relatively high level of variability (or when the measurement error is high). This is the preferred regime when it is possible to repeat the experimental runs and acquire new observations at the same locations. The newly observed trajectories are considered to be independent of each other.

  Note that when this kind of variability arises in calibration studies, it is advised to take the mean of the realizations and work on this single trajectory. We explore what happens when all trajectories are considered.

An illustration of these three regimes is given in fig. 3.8.

Definitions of consistency and asymptotic normality for estimators

Let the likelihood function be $\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\gamma})$, where $\mathbf{y}_{\mathrm{obs}}$ is the data of size n, and $\boldsymbol{\gamma}$ the vector parameter. Assume that there exists a "true value" $\boldsymbol{\gamma}^{(0)}$ so that the data were generated according to $\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\gamma}^{(0)})$.

An estimator $\hat{\boldsymbol{\gamma}}$ is said to be *consistent* if the estimator converges in probability to $\boldsymbol{\gamma}^{(0)}$:

$$\hat{\boldsymbol{\gamma}} \xrightarrow[n\to\infty]{P} \boldsymbol{\gamma}^{(0)} \qquad \Leftrightarrow \qquad \forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathrm{p}(|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{(0)}| > \varepsilon) = 0.$$

An estimator $\hat{\boldsymbol{\gamma}}$ is said to be *asymptotically normal* if $\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{(0)})$ converges in distribution to a normal distribution (multivariate if $\hat{\boldsymbol{\gamma}}$ is a vector):

$$\sqrt{n}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}^{(0)}) \xrightarrow[n\to\infty]{D} \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{V}).$$

We also define the *Fisher Information Matrix* to be the $\dim(\boldsymbol{\gamma}) \times \dim(\boldsymbol{\gamma})$ matrix $\mathbf{I}(\boldsymbol{\gamma})$ with (i, j) coefficient such that:

$$\mathbf{I}_{i,j} = \mathbb{E}_{\mathbf{y}_{\text{obs}}}\left[\left(\frac{\partial}{\partial\gamma_i}\log \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\gamma})\right) * \left(\frac{\partial}{\partial\gamma_j}\log \mathrm{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\gamma})\right)\right]. \tag{3.47}$$

Definitions of consistency and asymptotic normality for posterior distributions

Assume a prior distribution $\mathrm{p}(\boldsymbol{\gamma})$ over the model parameters. We propose definitions inspired from [Gelman, 2014, Chapter 4]. Mind their remark that these definitions are valid *almost surely* with respect to the true distribution of the data: notably, it is always possible to construct a sequence of data that will be unfavorable for the estimation. These problems are avoided by giving definitions valid only up to a set of measure zero.

The posterior distribution is *consistent* if, for all neighbourhoods A of $\boldsymbol{\gamma}^{(0)}$ with non-zero prior probability, we have:

$$\mathrm{p}(\boldsymbol{\gamma} \in A|\mathbf{y}_{\text{obs}}) \xrightarrow[n\to\infty]{} 1. \tag{3.48}$$

The posterior distribution is *asymptotically normal* with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$ if:

$$D_{\mathrm{KL}}(\mathrm{N}(\boldsymbol{\mu}, \mathbf{V})||\mathrm{p}(\boldsymbol{\gamma}|\mathbf{y}_{\text{obs}})) \xrightarrow[n\to\infty]{} 0. \tag{3.49}$$

Note that the notion of rate of convergence is present in the frequentist definition (the convergence must happen at the relatively fast rate of $1/\sqrt{n}$) but absent from the Bayesian definition.

Let us turn now to the theoretical results in the asymptotic regimes. We consider the likelihood function of the calibration problem of section 3.1.2, which is normal with mean and variance given in equations (3.4) and (3.5). In the increasing domain and repeated domain regimes, the existing results apply to the full likelihood function so that we have $\boldsymbol{\gamma} = (\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta})$ with previous notations. In the fixed domain regime, we only have results about the estimation of covariance structure (and not mean structure), so that we are restricted to the case where $\boldsymbol{\theta}$ is fixed, $\boldsymbol{\beta} = \mathbf{0}$, and $\boldsymbol{\gamma} = \boldsymbol{\psi}$.

Asymptotic results for ID

In this regime, the maximum likelihood estimator was proven to be consistent and asymptotically normal in [Mardia and Marshall, 1984]. There are some conditions required for these properties. The first one is the regularity of the likelihood function so that the Fisher Information Matrix exists (derivability of the likelihood) and is continuous w.r.t

γ. There are also requirements about the spacing between the observations in the design: if they are too close, the information they provide is redundant, and if they are too far away, we do not learn much about the correlation structure. We also require good identifiability of the mean structure and the covariance structure. All these conditions are discussed in more detail in [Bachoc, 2013a] chapter 4.

Some studies were also conducted when the design of observations is done in a specific shape: the most obvious is when the design is a grid with fixed spacing, as we have done in fig. 3.8. Then, more precise conditions for asymptotic normality can be derived, as in [Mardia and Marshall, 1984]. In [Bachoc, 2013a] such a design was considered with added random perturbations on the locations.

Asymptotic results for FD

In the fixed domain regime, estimation is more complex than in the previous regime. It is known, for example, that some hyperparameters can not be consistently estimated [Stein, 1999; Zhang, 2004], whereas they would be in the increasing domain regime. The notion of *microergodic* hyperparameters is central in this regime: it is proved that non-microergodic hyperparameters cannot be consistently estimated ([Bachoc, 2013a] chap. 4). Besides, there is no general result for hyperparameters that are microergodic: they might be, or not, consistently estimated. As a consequence, theoretical results are only derived for specific covariance families. We will give an illustration of microergodicity in the first numerical example of section 3.3.4.

In the case of the one-dimensional Matern kernel, it was proven that the hyperparameter $\sigma/l^{2\nu}$ is microergodic and that the Maximum Likelihood estimator is consistent ([Zhang, 2004]). In [Du et al., 2009], the estimator was also proven to be asymptotically normal. The addition of measurement error decreases the rate of convergence of the estimator, from $n^{-1/2}$ in the non-noisy case to $n^{-1/4}$ ([Chen et al., 2000]). It shows however that model discrepancy hyperparameters and measurement error can be jointly estimated in a consistent manner. The hyperparameters (σ, l), or σ alone, are both non-microergodic so they lack a consistent estimator ([Ying, 1991]).

For results that concern the Matern kernel in multiple dimensions (isotropic or tensorized), we refer to [Bachoc, 2013a]. A positive result for the squared exponential kernel is that the hyperparameter (σ, l) is microergodic ([Stein, 1999]), but for now only the maximum likelihood estimator for l is proven to be consistent [Loh and Lam, 2000].

Asymptotic results for RD

The specificity of the repeated domain asymptotic regime is that all observed trajectories are considered independent. Consequently, it does not suffer from problems related to microergodicity: looking at figure 3.7, it is intuitive that observing multiple trajectories at a low number of points in the domain is more favorable than observing one with a high number of points. We are brought back to the simple problem of estimating the mean vector and the covariance matrix using multiple observations drawn from a multivariate normal distribution and relating this estimation to the location parameters (for the mean

vector) and the covariance hyperparameters (for the covariance matrix). We give some necessary conditions that will be relevant in practice for estimation. More mathematical precision can be found in [Gelman, 2014, Appendix B].

Let $\boldsymbol{\gamma}^{(0)}$ be the minimizer of the KL divergence between the parametric family and the true distribution of the data. In the case where the true distribution belongs to the supposed family, naturally $\boldsymbol{\gamma}^{(0)}$ is the value used to generate the data. It is possible to prove that the posterior distribution is consistent and asymptotically normal towards $\mathrm{N}(\boldsymbol{\gamma}^{(0)}, (n\mathbf{I}(\boldsymbol{\gamma}^{(0)}))^{-1})$, assuming some regularity and identifiability conditions, notably:

- $\boldsymbol{\gamma}^{(0)}$ belongs to the interior of the search space $\Theta \times \Psi \times \mathbb{B}$,
- $\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\gamma})$ is differentiable at $\boldsymbol{\gamma}^{(0)}$,
- $\boldsymbol{\gamma}^{(0)}$ is the unique minimizer of the KL divergence.

Note that the second condition amounts to a regularity condition on $f_{\boldsymbol{\theta}}$ and $c_{\boldsymbol{\psi}}$, because if they are derivable the likelihood function is as well (section 3.2.2). The third condition also implies that the maximum likelihood is unique asymptotically.

We will now show how these properties relate to practical estimation following three numerical examples: In the first one, we perform hyperparameter estimation of a covariance structure in a non-microergodic context. In the second one, we realize a full calibration where the true process can be exactly reproduced by the computer model (*well-specified case*). We look at the posterior distributions when the number of observations increases, following the three asymptotic regimes. In the third example, we study the *misspecified case*, where the true process doesn't belong to the family of computer model predictions.

3.3.4 Numerical examples

Estimation of non-microergodic hyperparameters

As mentioned previously, with a Matern 1/2 kernel, the hyperparameters (σ, l) are not microergodic, but σ^2/l is so that only the latter can be estimated consistently in fixed domain asymptotics. We illustrate this property on figure 3.7, (a),(b),(c) by sampling trajectories for three values of σ^2, keeping the ratio σ^2/l constant. It can be seen that the local variations of trajectories in the three situations are identical, which is the reason why they are difficult to distinguish from each other.

We try a simple example where the observations are obtained from a trajectory of a one-dimensional Gaussian Process with kernel Matern 1/2 and hyperparameters $(\sigma, l) = (2, 4)$. Here, 400 non-noisy observations are obtained on a regular grid in the range $x \in [0, 5]$ (fig. 3.7 (d)). The hyperparameters (σ, l) are estimated, and the contours of the likelihood function are plotted (3.7 (e)). The likelihood has the shape of an ellipse perfectly aligned with the isoline $\log \sigma = 2 \log l$ (dashed line), and the generating value of hyperparameters lies on that isoline, which shows that the ratio σ^2/l is correctly estimated. The maximum likelihood estimator is $(\hat{\sigma}, \hat{l}) = (3.57, 12.88)$, which is not

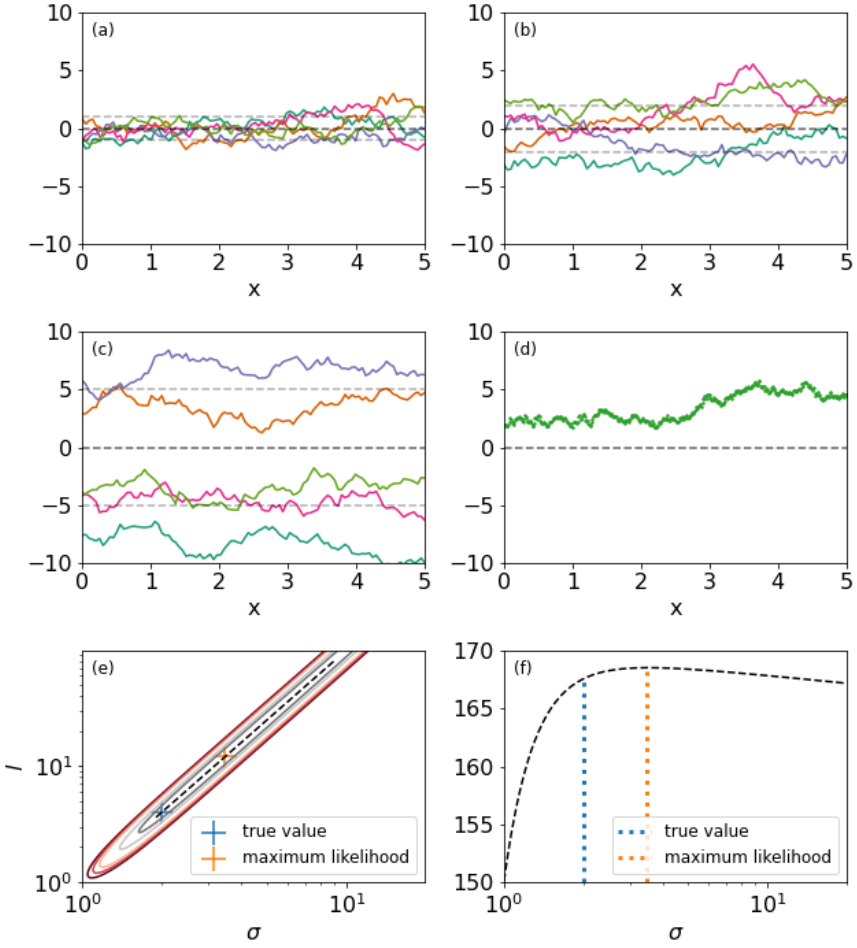


Figure 3.7: Estimation of non-microergodic hyperparameters. (a),(b),(c): samples from Matern 1/2 kernel with (σ,l)=(1,1),(2,4) and (5,25). (d): 400 observations of a trajectory from kernel (b). Could they have come from kernel (c)? (e): contourplot of the log-likelihood depending on σ and l. The distance between two continuous lines is 2.5 in log-probability. The dashed line follows $\sigma^2/l = 1$ which is an isoline of the microergodic hyperparameter. (f): values of the log-likelihood along the dashed line.

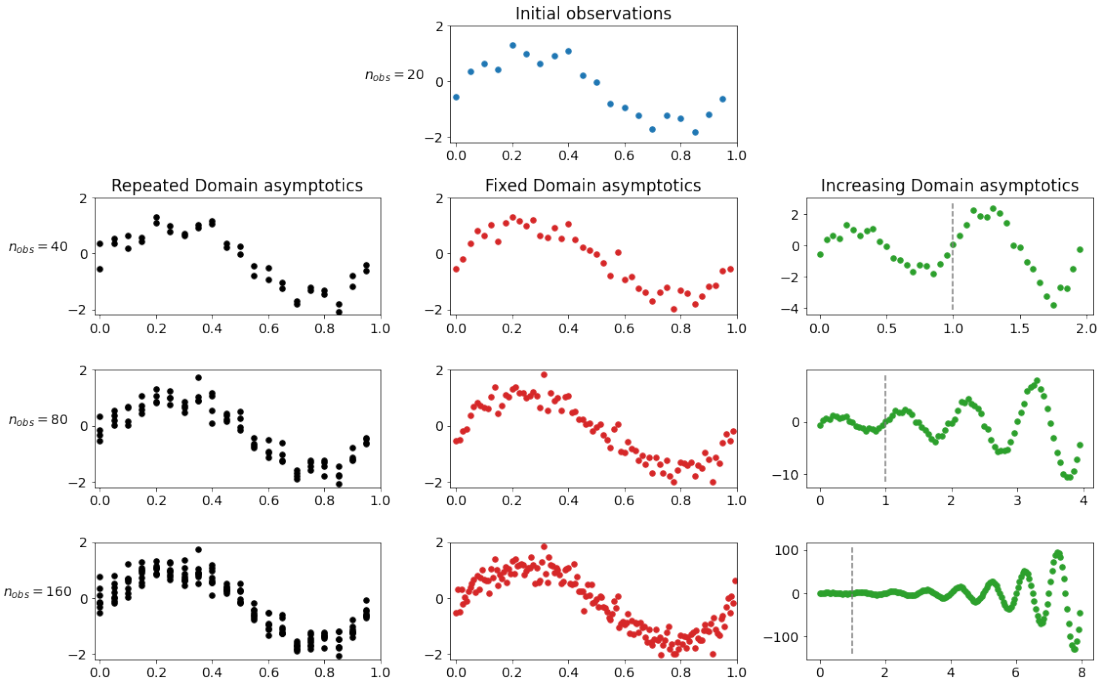


Figure 3.8: Asymptotic regimes: Repeated Domain (left column), Fixed Domain (center column) and Increasing Domain (right column). Going from top to bottom shows an increase of observations, and within each regime, the design of each step includes one of the previous steps. All designs are generated from the initial observations (blue, at the top). A vertical line delimits the initial domain in the Increasing Domain regime.

close to the generating values. Plotting the values of the likelihood function over the isoline (3.7 (f)) reveals that values up to $\sigma = 10$ are considered more likely than the generating value.

Looking at the figures, we can instantly guess that the observations are unlikely to come from the kernel $\sigma = 1$ because they are too far from zero. It is harder to distinguish between the values $\sigma = 2$ or $\sigma = 5$. This fact is seen in fig. 3.7 (f) as the log-likelihood decreases sharply for low values of σ, but the decrease for high values is way slighter.

In this example, the estimation of non-microergodic hyperparameters leads to a flat likelihood function in some direction, so the maximum likelihood estimator is hardly reliable. As a silver lining, we remark that it is not dramatic for our calibration problem because we do not care about the value of hyperparameters, rather the distributions of $z_{\boldsymbol{\theta}}$ that stem from it. Similar distributions of $z_{\boldsymbol{\theta}}$ will lead to similar conclusions on $\boldsymbol{\theta}$, which is the main goal of calibration.

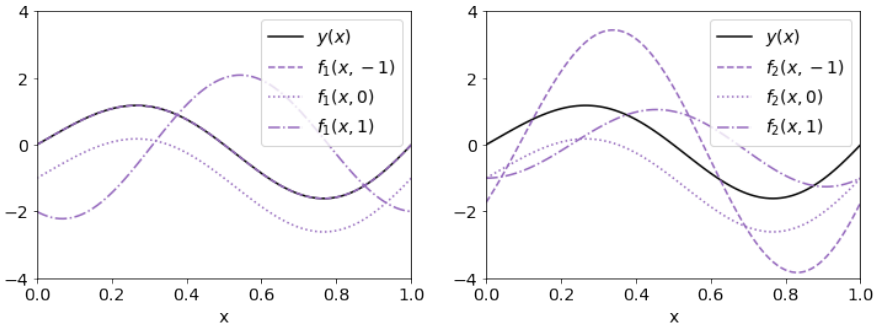


Figure 3.9: True process and computer models in the numerical example of Tuo and Wu. The model f_1 is well-specified and f_2 is misspecified.

Calibration with a well-specified model

We turn towards a calibration problem by borrowing the example from [Tuo and Wu, 2015b] where the true process is :

$$y(x) = \exp(2\pi x/10)\sin(2\pi x), \tag{3.50}$$

we consider an initial number of 20 observations, regularly spaced in the domain $x \in [0, 1]$. The measurement error is $\varepsilon(x) \sim \mathrm{N}(0, \sigma^2_{\mathrm{mes}})$, where $\sigma^2_{\mathrm{mes}} = 0.1$ is known a priori. The computer model is given by the function:

$$f_1(x, \boldsymbol{\theta}) = y(x) - |\theta + 1|(\sin(2\pi\theta x) + \cos(2\pi\theta x)). \tag{3.51}$$

Note that the model verifies $y(x) = f_1(x, -1)$ so that there exists a value of model parameters for which the true process can be reproduced. This is the *well-specified* context. Some model predictions are plotted on figure 3.9 (left). We consider the range of parameters such that $\theta \in [-2, 2]$. For the model discrepancy, we choose mean zero and the squared exponential kernel

$$c_\psi(d) = \sigma^2 \exp(-d^2/2l), \tag{3.52}$$

where the hyperparameters σ and l are learned, so we are in the microergodic regime. We perform the calibration first using the 20 observations, then with grids of increasing size comprising 40, 80 and 160 observations. We consider three variants of grids according to the three asymptotic regimes, RD, FD and ID (see fig. 3.8). The posterior distributions are sampled with the M-H algorithm with 10^6 steps with 10% of burn-in. After running the chain, the last 5×10^4 samples are taken as samples of the posterior. Visual checks in the form of autocorrelation plots were performed to ensure good mixing of the chains.

The posterior distributions in the asymptotic limits are plotted on figure 3.10. In all three regimes, the posterior of θ seems to converge to a Dirac centered around the generative value $\theta = -1$. The variance decreases with the number of observations (not shown here), but no definite speed of convergence was observed, so we cannot positively

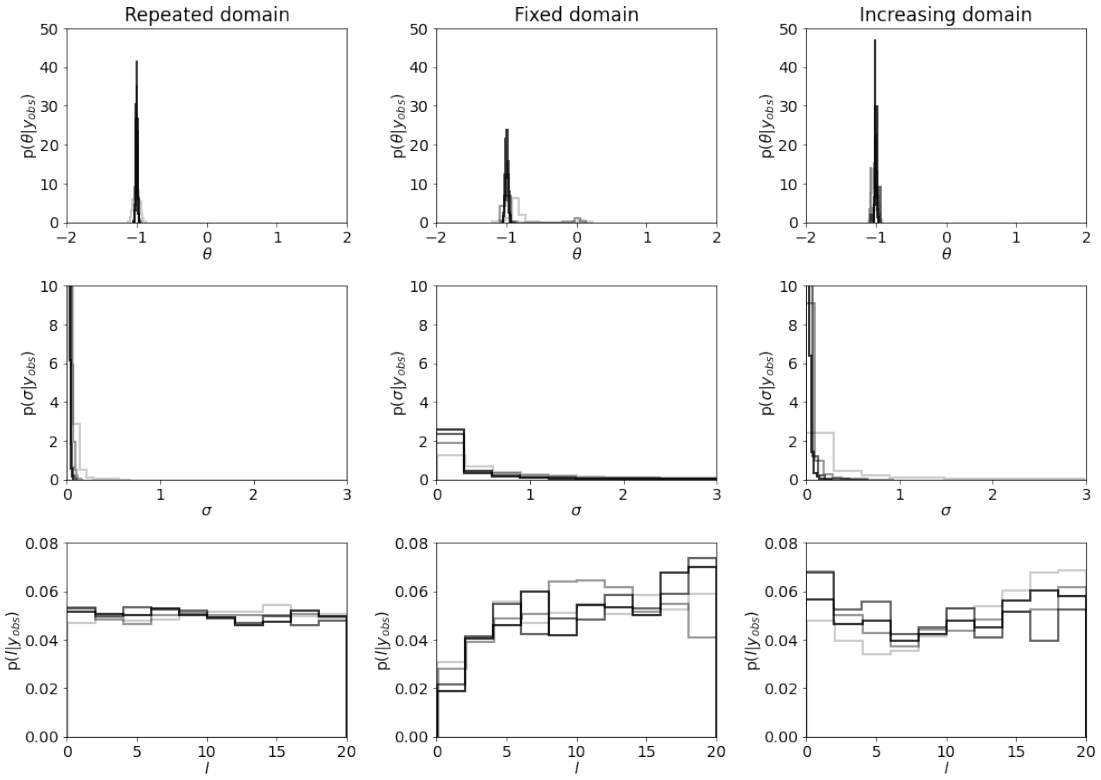


Figure 3.10: Posterior distributions in the well-specified case. The lines become darker as the number of observations in the grid is increased. They correspond to $n_{obs} = 100, 300, 600, 900$.

conclude that it converges to zero. The distributions of σ in each regime concentrate to the value 0, which is the true value in the well-specified case. When the variance is low, the correlation length bears no influence on the likelihood function, which explains the flat posterior distributions for l in the three regimes. Note that, in general, this configuration presents a challenge for MCMC algorithms (mixing at the boundary + one non-influential hyperparameter), but still, we were able to obtain independent samples.

These plots also reveal a significant difference in the three asymptotic regimes: in RD and ID, the true value of σ is narrowly identified, whereas it is less so in the FD regime. The ambiguity around the value $\theta = 0$ is still present in the RD regime for $n_{\text{obs}} = 100, 300$, but eventually the three distributions of θ are able to identify the true value of the parameter.

We learn from this example that, in the well-specified case, with a high number of observations, the calibration of the computer model does not lead to a normal joint posterior because the true hyperparameters lie on the boundary of the search space. On the positive side, this doesn't perturb parameter estimation because the true value of parameters is quickly identified with a high degree of certainty.

Calibration with a misspecified model

Following the second example of [Tuo and Wu, 2015b], we now take the computer model to be:

$$f_2(x, \boldsymbol{\theta}) = y(x) - (\sqrt{\theta^2 - \theta + 1}) * (\sin(2\pi\theta x) + \cos(2\pi\theta x)), \tag{3.53}$$

while keeping the true process and the observations from the previous example. Here the model is misspecified, meaning there doesn't exist a value of $\boldsymbol{\theta}$ for which it is equal to the true process. Some model predictions are plotted on figure 3.9 (right subfigure).

Histograms of the posterior densities are plotted on figure 3.11. As in the previous example, the posterior distributions for θ converge to a single value when the number of observations increases. The posteriors of σ are concentrated around 1, with a bigger dispersion than in the well-specified case. They also feature a heavy right tail. For the correlation length l, the distribution in the FD regime is flat, and for the other two regimes, it becomes concentrated at the upper bound $l = 20$. Note that the initial domain is $x \in [0, 1]$ and the domain in the ID regime with $n_{obs} = 900$ is $x \in [0, 45]$. This behaviour occurs because the inferred model discrepancy is constant across the domain, which is confirmed by the shape of the model predictions at $\theta = 0$ (see fig. 3.9). Thus, the largest correlation length is always preferred because it increases the value of the likelihood function.

Once again, the parameter posteriors are consistent within each regime. Posteriors of each regime agree with each other, and the posterior in the FD regimes shows more variance than in RD and ID.

In this case, there is no value of θ that plays the role of a "true value", yet in the three asymptotic regimes, the posteriors are consistent. This result is reassuring for the

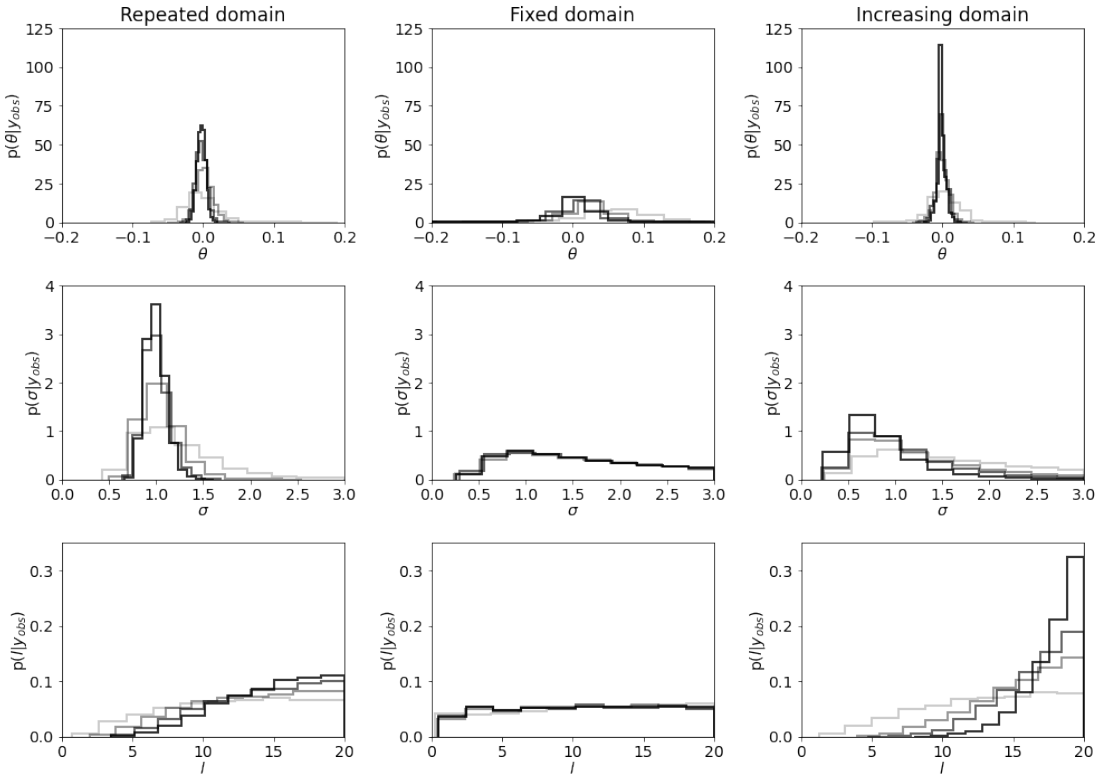


Figure 3.11: Posterior distributions in the misspecified case. The lines become darker as the number of observations in the grid is increased. They correspond to $n_{obs} = 100, 300, 600, 900$.

use of Bayesian techniques in calibration. In [Tuo and Wu, 2015a] it is shown that, when the covariance structure is fixed (KOH estimation), in the FD regime, the parameter posterior will be consistent towards the value that minimizes the Reproducing Kernel Hilbert Space norm induced by the covariance structure. No result of this kind exists for random covariance structures.

3.4 Conclusion

In this chapter, we have developed a new interpretation of the model error in the Bayesian calibration context. From this followed a statistical framework with an adaptive model discrepancy term. This framework was already considered in [Higdon, Kennedy, et al., 2004; Higdon, Gattiker, et al., 2008] as the Full Bayesian Approach, as well as in [Plumlee, 2017], but with different treatments of model parameters. These considerations have led to suppose a functional relationship between model parameters and model discrepancy hyperparameters, which brought us to the FMP estimation, a cost-effective approximation of the posterior distribution under two assumptions that are related to the good choice of kernel family. Continuity properties were proven, as well as gradients of the criterion to optimise, for an efficient numerical resolution. By approximating the joint posterior as a Gaussian mixture with well-separated modes, we have proven the accuracy of the FMP method over the KOH method, as it does not fall prey to the false confidence fallacy and does not miss probability modes. In case the modes are not well-separated, we did not find theoretical results, but the FMP method also seems promising.

Two numerical studies complete this chapter. In the first one, we focus on a problem where the hyperparameters are non-microergodic and show that, even with a sufficient number of observations, the likelihood function is approximately flat, thus maximum likelihood estimation of kernel hyperparameters can be compromised, but that it does not influence the calibrated values of model parameters. The second study concerns two cases, a well-specified model that can reproduce the "reality" and a misspecified model that can't. They are both treated when the number of observations tends to infinity under three asymptotic frameworks. Note that the RD asymptotic frameworks that we introduced are studied very little in calibration problems, although it exhibits good asymptotic properties. The Gaussian approximation proves inadequate for the joint posterior but the parameter posteriors are consistent.

Note that other types of functional relationships could be encoded instead of the FMP estimation that we have chosen. For example, one could use cross-validation estimates depending on the model parameter value, something that we expect to be less informative but more robust to model misspecification. Another idea is to integrate prior knowledge about the model discrepancy via this relationship: see for example [DeVolder et al., 2002, Section 3], in which is discussed that it is possible, for different values of the input parameters, we can expect varying degrees of accuracy in the solution (in his case, numerical accuracy), which would correspond to a different prior distribution for the model discrepancy term.

Chapter 4

Efficient calibration of computer models

Contents

In this chapter, we introduce numerical techniques, some novel and some established, to perform calibration efficiently. The first section deals with MCMC sampling, in which we detail its application to calibration problems. An Importance Resampling algorithm is given to improve the quality of the FMP posterior sample at the cost of reducing its size. The second section deals with surrogate modelling in calibration, either directly on the computer code when its output is large (PCA technique) or on statistical quantities of the FMP method itself. A third section presents an algorithm to adequately place training points for the FMP surrogates in regions that are plausible a posteriori by iteratively sampling from MCMCs and selecting points in the sample. All the proposed methods are illustrated in numerical examples along the sections.

4.1 Sampling the posterior with MCMC algorithms

The bread and butter of Bayesian numerics is the estimation of integrals, more precisely those of the form:

$$C = \mathbb{E}_{\boldsymbol{\theta}}\left[f(\boldsymbol{\theta})\right] = \int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}, \tag{4.1}$$

where $\boldsymbol{\theta}$ is a random variable with distribution $\mathrm{p}(\boldsymbol{\theta})$, and f is a function of $\boldsymbol{\theta}$.

This task can often be performed using some quadrature rules, which provide two things: a set of points over which the target function and/or its derivatives are evaluated, and a linear formula to recombine these evaluations into an estimator of C. In this family of methods, we find, for instance, the midpoint rule, the trapezoidal rule, or the Gauss quadratures in which the target function is approximated as a polynomial. Quadrature rules are a convenient way of estimating integrals. Still, the upper error bound on the approximation depends exponentially on the dimension of $\boldsymbol{\theta}$, so in high-dimensional problems, it explodes (which in our case is the dimension of model parameters plus all hyperparameters of the model discrepancy). This behaviour is referred to as the *curse of dimensionality*, and it is the reason why we turn to techniques involving random sampling, namely Monte-Carlo methods.

4.1.1 Estimation of integrals using Monte-Carlo

We present the classical Monte-Carlo estimator for the quantity C. It relies on the fact that we can draw correctly a i.i.d. sample of the probability distribution p, that we note $\{\boldsymbol{\theta}_i\}_{i<n}$. The following subsections focus on how a good sample can be obtained but assume we have it now. Thus, the Monte-Carlo estimator is given by the empirical mean of the sample:

$$\hat{f} = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i). \tag{4.2}$$

This estimator is unbiased because $\mathbb{E}[\hat{f}] = C$. Its variance is easily obtained by using the property that the sample is i.i.d:

$$\begin{aligned}
\mathrm{Var}[\hat{f}] &= \mathrm{Var}[\frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta}_i)] \\
&= \frac{1}{n^2}\mathrm{Var}[\sum_{i=1}^{n} f(\boldsymbol{\theta}_i)] \\
&= \frac{1}{n}\mathrm{Var}[f(\boldsymbol{\theta})] \\
&= \frac{1}{n}\left[\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})^2\mathrm{p}(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta} - \left(\int_{\boldsymbol{\theta}} f(\boldsymbol{\theta})\mathrm{p}(\boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}\right)^2\right].
\end{aligned} \tag{4.3}$$

Thus the variance of the estimator depends only on the function variance (over which we have no control), and the sample size. This variance is often estimated using the in-sample variance:

$$\mathrm{Var}[f(\boldsymbol{\theta})] \approx \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(f(\boldsymbol{\theta}_i) - \hat{f})^2. \tag{4.4}$$

According to the Central limit theorem, $\hat{f}$ is asymptotically normal. We can then write the following 95% confidence interval for C:

$$I_{C,95\%} = [\hat{f} - 1.96\frac{\sqrt{\text{Var}[f(\boldsymbol{\theta})]}}{\sqrt{n}}, \hat{f} + 1.96\frac{\sqrt{\text{Var}[f(\boldsymbol{\theta})]}}{\sqrt{n}}], \tag{4.5}$$

which can be computed by replacing the target variance with its estimator:

$$\hat{I}_{C,95\%} = [\hat{f} - 1.96\frac{\hat{\sigma}}{\sqrt{n}}, \hat{f} + 1.96\frac{\hat{\sigma}}{\sqrt{n}}]. \tag{4.6}$$

Thus we find the famous error rate in $\frac{1}{\sqrt{n}}$. This means that if one wishes to increase the precision of the estimator by a factor 10, we should get a sample that is 100 times bigger. Some numerical techniques improve the Monte-Carlo estimator by getting samples from distributions other than $\text{p}(\boldsymbol{\theta})$. In this family of methods called variance reduction techniques, we find importance sampling or stratified sampling [Gelman, 2014].

The targeted variance might be high with a high dimension of $\boldsymbol{\theta}$ so that a significant sample size would be required - but it's still better than suffering from the curse of dimensionality. The catch in the Monte-Carlo techniques is that it is not easy to get an independent sample from a high dimensional probability density. Besides simple distributions such as the multivariate Gaussian, this task is arduous because we have no way of knowing where the density is high without evaluating it on the whole domain, and if we wanted a fine grid estimation of the density, we would fall back to the curse of dimensionality. If the shape of the density is disadvantageous, for example, it presents multiple narrow peaks, large flat zones, or both; this introduces additional difficulty. Fortunately, there is a wide assortment of sampling techniques that belong in the MCMC family [Robert and Casella, 2010], and in the next section, we present the classical M-H algorithm.

Application to the calibration problem

In the calibration problem, the distribution that was noted $\text{p}(\boldsymbol{\theta})$ plays the role of the posterior distribution $\text{p}(\boldsymbol{\theta}, \boldsymbol{\psi}, \boldsymbol{\beta}|\mathbf{y}_{\text{obs}})$, so MCMC algorithms are employed to obtain an independent sample of the posterior. After obtaining this sample, the whole posterior distribution can be estimated using a histogram or kernel density estimation techniques. The Monte-Carlo estimator intervenes if one wants to compute the mean or variance of the posterior, the predictive density (eqs. (3.15) and (3.16)), or any function of the model parameters that could be of interest for a given application, such as quantiles for rare event estimation, or failure probabilities.

4.1.2 The Metropolis-Hastings algorithm

Assume that we wish to draw samples from the target distribution $\text{p}(\boldsymbol{\theta})$, but it is so complex that we are only able to evaluate an unnormalized version of it, $\text{p}^*(\boldsymbol{\theta}) = \text{p}(\boldsymbol{\theta})/Z$ with unknown Z. This is typically the case in a calibration setting because following Bayes' theorem, the posterior distribution is known up to a multiplicative constant. The

privileged solution is to employ the M-H algorithm [MacKay, 2003], which simulates a random walk in the parameter space, to explore the whole posterior distribution correctly. Rejection Sampling is an alternative, but it requires the delicate formulation of a proposal distribution. In the family of Monte-Carlo methods, Gibbs sampling or Slice sampling [Gelman, 2014] are simpler methods that require more knowledge about the target distribution. More involved techniques can also be envisioned if the mixing obtained by M-H is insufficient due to the complexity of the target distribution, such as Annealed Importance Sampling, Reversible Jump Metropolis-Hastings, or adaptive MCMC [Andrieu and Thoms, 2008].

The M-H algorithm constitutes a Markov Chain in which a series of states $\{\boldsymbol{\theta}^{(i)}\}_{i<n}$ is explored, following random rules of succession. The *Markov property* specifies that the choice of the following state depends only on the current state and not on the history of visited states. If the random rules are properly tuned, the series of explored states (or a subsequence) will look like an independent sample of the target distribution. A necessary condition to get such a sample is that the target distribution is *stationary* with respect to the chain. By definition, a distribution is stationary if it verifies the *global balance equation*, which in the case of a discrete sample space write:

$$\mathrm{p}(\boldsymbol{\theta}_j) = \sum_{\boldsymbol{\theta}_i} \mathrm{p}(\boldsymbol{\theta}_i)\mathrm{p}(\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j), \tag{4.7}$$

where $\mathrm{p}(\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j)$ is the probability to transition from state i to state j. A stronger condition is given by the *local balance equations*:

$$\mathrm{p}(\boldsymbol{\theta}_j)\mathrm{p}(\boldsymbol{\theta}_j \to \boldsymbol{\theta}_i) = \mathrm{p}(\boldsymbol{\theta}_i)\mathrm{p}(\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j), \quad \forall i, j. \tag{4.8}$$

It is easy to check that local balance implies global balance, so the target distribution is stationary.

The specificity of the M-H is to use a two-step rule of succession: first, a candidate state is proposed; second, it can be either accepted or rejected. If it is accepted, it becomes the current state. Otherwise, the current state remains, and a new candidate state is proposed. The user has to specify a *proposal distribution* $\mathrm{q}(\boldsymbol{\theta}_j|\boldsymbol{\theta}_i)$, which is the probability to propose the state j if the current state is i. The candidate point is then accepted with probability α, given by:

$$\alpha = \min\left(\frac{\mathrm{p}^*(\boldsymbol{\theta}_j)\mathrm{q}(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)}{\mathrm{p}^*(\boldsymbol{\theta}_i)\mathrm{q}(\boldsymbol{\theta}_j|\boldsymbol{\theta}_i)}, 1\right). \tag{4.9}$$

It follows that, in the M-H algorithm, we have $\mathrm{p}(\boldsymbol{\theta}_i \to \boldsymbol{\theta}_j) = \mathrm{q}(\boldsymbol{\theta}_j|\boldsymbol{\theta}_i) * \alpha$, and a simple calculation allows to check that the target distribution verifies the local balance equations and is consequently stationary.

The stationarity of distribution does not necessarily imply that the chain is convergent towards this distribution and that the visited states will resemble a sample of it. Other technical conditions are required, such as reducibility, invariance, and aperiodicity

[Robert and Casella, 2010]. We consider that these are all verified in our algorithms and applications.

The usual choice for the proposal distribution is a Gaussian distribution, centered at the current state. Its covariance matrix $\mathbf{C}_{\text{prop}}$ can be user-specified, and we obtain the procedure summarized in algorithm 1.

Algorithm 1 Metropolis-Hastings algorithm with Gaussian proposal.

Specifiy an initial state $\boldsymbol{\theta}^{(0)}$ and a proposal covariance $\mathbf{C}_{\text{prop}}$.
for $n < n_{\max}$ **do**
  Draw a candidate point $\boldsymbol{\theta}^{\text{cand}} \sim N(\boldsymbol{\theta}^{(n)}, \mathbf{C}_{\text{prop}})$.
  Draw $U \sim U(0,1)$.
  Accept $\boldsymbol{\theta}^{\text{cand}}$ with probability $\max(\frac{\text{p}(\boldsymbol{\theta}^{\text{cand}}|\mathbf{y}_{\text{obs}})}{\text{p}(\boldsymbol{\theta}^{(n)}|\mathbf{y}_{\text{obs}})}, U)$.
end for

This version of the M-H algorithm has two undesired properties. One is that the first states are dependent on the initial state $\boldsymbol{\theta}^{(0)}$, which is user-defined. Besides, the quality of the exploration relies on the good specification of the proposal covariance $\mathbf{C}_{\text{prop}}$, which should be close to the covariance matrix of the target distribution, of which we have little information a priori. The common approach to correct these properties is to define the first steps of the chain to be the *burn-in* phase, which will not be considered a part of the chain when extracting samples. The proposal covariance matrix can also be estimated from the states of the burn-in phase, and according to a normality argument [Andrieu and Thoms, 2008] we take:

$$\mathbf{C}_{\text{prop}} = \frac{2.38^2}{d}\mathbf{C}_{\text{burn}}, \tag{4.10}$$

where $\mathbf{C}_{\text{burn}}$ is the empirical covariance of the burn-in states, and d is the dimension of the chain. A common choice for the length of the burn-in phase is $n_{\text{burn}}/n = 10\%$.

Forming the sample

A thorny question among MCMC practitioners is to find the best way to extract an independent sample from the runs of the chain. Also, should one long chain be preferable to several small chains to get independence? In the following paragraphs, we discuss multiple techniques for forming the sample in a discussion inspired from [Gamerman and Lopes, 2006].

Assume that we wish to obtain a sample of size s. One idea is to run s chains with independent initial states, ensure that they are converged, and take the final state of each to constitute the posterior sample. For good independence properties, the initial states should have a large dispersion. This technique is arguably the most expensive to get an independent sample.

Another idea is to extract the sample from a single long chain. After the chain is converged by the stationarity property, any sequence of states of size s is a sample of the

posterior distribution. The catch is that the draws are not independent of each other due to being successive states. This is not a problem if the self-correlation of the chain is small enough with respect to the sample size so that it excludes some undesirable effects such as insufficient exploration or too few transitions between probability zones. The self-correlation of the chain is of utmost importance for convergence diagnostics, as will be discussed in section 4.1.4.

A method to obtain quasi-independence of draws using a single chain would be not to consider s successive states, but rather s states taken every k step of the chain. If the auto-correlation lag is small with respect to k, then the sample can be considered independent. It is crucial to recognize that, in terms of the precision of the Monte-Carlo estimators, there is no benefit in doing so compared to using all of the chain states. This method is advantageous when it would be counterproductive to use all chain states (due to computer storage or calculation time when using the sample).

It can be argued that running a small number m (single-digit) of chains with dispersed initial states is beneficial since it allows to diagnose undesired behaviours of specific chains and average them out. Then, to build a sample, one can take s/m consecutive states from each chain, or in the spirit of the previous technique, take states every k step of each chain.

In conclusion, running a few independent chains is often a good idea to lower the probability of bad single-chain behaviour. The mutual behaviour of the chains can also be an indicator of convergence, as we will discuss in section 4.1.4: if they all quickly settle in one region, then it suggests that running one chain is enough.

Application to calibration

In the calibration problem, the M-H algorithm is used to get a sample from the posterior distributions. It is naturally necessary to run long chains and to ensure that they are well-mixed, yet we sometimes do not wish to retain all states as a sample because of the numerical cost required to use them in prediction. Thus, we used regularly-separated states to form the sample in this work. Note also that the majority of the computer cost in the method comes from the computer model evaluations so that once these are done, and the surrogate model is built, it does not take long to try multiple configurations for the chains.

The MCMC algorithms present some specificities depending on the calibration method chosen (KOH, FMP, or Bayes). We detail them below:

- For the KOH method, the target distribution is $\mathrm{p}_{KOH}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) \propto \mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \hat{\boldsymbol{\psi}}_{KOH})$. We have precedently noted that the main cost of evaluating the likelihood function lies in the inversion of the matrix $\mathbf{V}_{\boldsymbol{\psi}}$. Because the hyperparameters are fixed, the inversion needs to be done once and then stored. The MCMC required for the KOH method is the cheapest of the three methods.
- For the FMP method, the target distribution is $\mathrm{p}_{FMP}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}}) \propto$

$\mathrm{p}(\boldsymbol{\theta})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}))$. The optimal hyperparameters must be evaluated for each candidate point, and the matrix inversion must be performed. We replace the most costly step, the optimisation, with a surrogate evaluation, but the cost remains higher than the KOH MCMC. Since the optimal hyperparameters are known for each point of the sample, it is possible to apply at virtually no cost an importance resampling algorithm to improve the FMP sample (see next section).

- For the Bayes method, the target distribution is $\mathrm{p}_{Bayes}(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}_{\mathrm{obs}}) \propto \mathrm{p}(\boldsymbol{\theta}, \boldsymbol{\psi})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi})$. Here the matrix inversion must be done at each step, and the dimensionality of the MCMC is incremented of the dimension of $\boldsymbol{\psi}$ (and eventually $\boldsymbol{\beta}$) so that it is harder to reach good mixing. Besides, exploring in the dimension of $\boldsymbol{\psi}$ might present additional difficulties such as exploring near the boundaries of the domain (see σ in the numerical examples of section 3.3.4), or hyperparameters that are not influent (see l in the same examples). As a consequence, the Bayes MCMC requires more samples for equivalent quality of mixing and is the most expensive method of the three.

4.1.3 Importance Resampling of the FMP sample

We present a technique to get a sample of the Bayesian posterior without resorting to a MCMC algorithm on the full joint space, using the sample of the FMP approximation obtained with the M-H algorithm of the previous section. Let $\{\boldsymbol{\theta}^{(i)}\}_{i<n_s}$ be this sample of size n_s. This is the Importance Resampling algorithm [Gelman, 2014], in which we compute for each point in the sample an importance weight $\omega(\boldsymbol{\theta}^{(i)})$. Then, a new sample of size n'_s is created by drawing from the old sample, with weights $\{\omega(\boldsymbol{\theta}^{(i)})\}_{i<n_s}$. Naturally, the size of the new sample should be reduced by the procedure, and the authors advocate for a ratio $n_{s'}/n_s = 1/20$.

The importance weights generally require to evaluate the target distribution on the sample, but in this case it can be avoided with the following approximation:

$$\begin{aligned}
\omega(\boldsymbol{\theta}^{(i)}) &\propto \frac{\mathrm{p}_{\mathrm{Bayes}}(\boldsymbol{\theta}^{(i)}|\mathbf{y}_{\mathrm{obs}})}{\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}^{(i)}|\mathbf{y}_{\mathrm{obs}})} \\
&\propto \frac{\mathrm{p}(\boldsymbol{\theta}^{(i)}) \int_{\boldsymbol{\theta}'} \mathrm{p}(\boldsymbol{\theta}'|\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}'))\,\mathrm{d}\boldsymbol{\theta}'.}{\mathrm{p}(\boldsymbol{\theta}^{(i)})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}^{(i)}))} \\
&\propto \frac{\int_{\boldsymbol{\theta}'} \mathrm{p}(\boldsymbol{\theta}'|\mathbf{y}_{\mathrm{obs}})\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}'))\,\mathrm{d}\boldsymbol{\theta}'.}{\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}^{(i)}))} \\
&\propto\!\!\!\!\sim \frac{(1/n_s)\sum_{j<n_s} \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}^{(j)}))}{\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}^{(i)}, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}^{(i)}))},
\end{aligned} \tag{4.11}$$

where we have used an intermediate result of equation (3.13) obtained from the FMP approximation, and also the fact that the FMP sample is close to a posterior sample.

In [Gelman, 2014], the sampling technique is also discussed, pondering between sampling with replacement or without. They argue that if the weights are unevenly distributed (for example, a lot of small weights and a few big ones), then if there is a

replacement, the same values of $\boldsymbol{\theta}$ will be drawn in repetition. Thus, they advocate for not reusing samples, even if better statistical properties are achieved in the context of replacement. They also propose to use Pareto smoothing with replacement [Vehtari et al., 2021], where the most extreme weights are reduced using a generalized Pareto distribution fit.

The benefit of using this procedure in the context of calibration is that by computing importance weights, we can compensate the volume effects that are not seen by the FMP method when the posterior is multimodal (section 3.3.2). This correction will be demonstrated numerically in the example of section 4.1.5. Besides, a crucial condition for the Importance Resampling algorithm to function correctly is that the approximating distribution should cover all the probability regions of the target distribution. We are confident the FMP approximation can achieve it because of the conservativity of the approach demonstrated in the case of the multimodal posterior.

4.1.4 Diagnostics for convergence of MCMC

There is a world between the theoretical properties of a chain, which are valid when the number of steps approaches infinity, and the practical consideration that a sample, extracted from a given chain with a finite number of steps, can be considered independently drawn from the limit distribution of the chain. Thus, multiple techniques try to assess whether or not a given series of states can be considered to have converged, so it is safe to form a sample from it. Using such statistical techniques, the convergence of the chain can never be proved, only corroborated. Thus, one must always treat such estimates with suspicion.

On a single multidimensional chain

Suppose that we have ran a multidimensional chain where $\{\boldsymbol{\theta}^{(i)}\}_{i<n_s}$ is the sequence of visited states, $\theta_j^{(i)}$ the j-th coordinate of the i-th point, and $\overline{\boldsymbol{\theta}} = (1/n_s)\sum_{i<n_s}\boldsymbol{\theta}^{(i)}$ the global mean. The mixing of the chain can be assessed by computing the self-correlation of the chain states in each dimension, noted c_j. It is a function of the lag l, with expression given by:

$$c_j(l) = \frac{\sum_{i<n_s-l}(\theta_j^{(i)} - \bar{\theta}_j)(\theta_j^{(i+l)} - \bar{\theta}_j)}{\sum_{i<n_s-l}(\theta_j^{(i)} - \bar{\theta}_j)^2}. \tag{4.12}$$

An estimator for the self-correlation distance τ_j is given by

$$\tau_j = \int_{l=0}^{l_{\max}} c_j(l)\,\mathrm{d}l, \tag{4.13}$$

where $l_{\max}$ should be chosen large enough to capture the decrease of the self-correlation to zero. The self-correlation distance can be estimated with other quantities, such as the first time at which c_j crosses zero, or the inverse of the derivative of c_j at $l = 0$. We define the total correlation distance $\tau = \max_j \tau_j$, which we will use later to compare the quality of the mixing of the chains between calibration techniques.

There are more advanced techniques available to study single-chain states. One approach is to fit time series models on the sequence of states [Geweke, 1991] and study convergence by computing the ergodic mean of the states. In batch analysis [Flegal and Jones, 2010], the states of a single chain are divided into batches, overlapping or not, and their properties are compared relative to each other. A more detailed review of MCMC diagnostics can be found in [Roy, 2019].

On multiple chains

The idea of using multiple chains is to compare individual and global behaviours. More specifically, the idea presented in [Gelman and Rubin, 1992] is to run multiple chains in parallel with overdispersed initial states, estimate the dispersion in these chains, and check how it compares to the dispersion between chains. Assume that there are m chains with length n, and note $\boldsymbol{\theta}_{[j]}^{(i)}$ the i-th state of the j-th chain, with $i \leq n$ and $j \leq m$. Thus, we can write the variances between chains B and within chains W as:

$$B = \frac{n}{m-1} \sum_{j \leq m} (\bar{\boldsymbol{\theta}}_{[j]} - \bar{\boldsymbol{\theta}})^2 \quad \text{and} \quad \frac{1}{m(n-1)} \sum_{j \leq m} \sum_{i \leq n} (\boldsymbol{\theta}_{[j]}^{(i)} - \bar{\boldsymbol{\theta}}_{[j]})^2, \tag{4.14}$$

where $\bar{\boldsymbol{\theta}}_{[j]}$ is the average of states of the j-th chain, and $\bar{\boldsymbol{\theta}}$ is the average of these averages. A consistent estimator for the target variance is given by $\hat{\sigma}^2 = (1-1/n)W + (1/n)B$. The idea is that, before the chains are converged, W will underestimate the target variance because the parameter space will not have been explored thoroughly. The estimator $\hat{\sigma}^2$ will on the contrary underestimate the target variance because the starting points of the chains are overdispersed. Thus, the estimator of potential reduction is given by:

$$\hat{R} = \sqrt{\frac{\hat{\sigma}^2}{W}}, \tag{4.15}$$

and convergence of the chains can be assumed if $\hat{R}$ is close enough to 1, with the value of 1.2 being suggested by the authors.

4.1.5 Numerical example: calibration of a sensitive model

Let us compare the three calibration techniques, Bayes, KOH and FMP on a simple problem where the predictions of the model are sensitive to the parameter values.

Assume that the true process and the computer model are given by:

$$y(x) = x, \qquad f(x, \theta) = x * \sin(2x\theta) + (x + 0.15) * (1 - \theta). \tag{4.16}$$

The range of variations of variables are $(x, \theta) \in [0, 1] \times [-0.5, 1.5]$. The calibration is performed using 8 observations with low Gaussian noise $\sigma_{\text{mes}} = 10^{-2}$. The observations and some model predictions are plotted on figure 4.1.

The prior for the model discrepancy is a zero mean GP with squared exponential kernel:

$$c_{\boldsymbol{\psi}}(d) = \sigma^2 \exp(-\frac{d^2}{2l^2}). \tag{4.17}$$

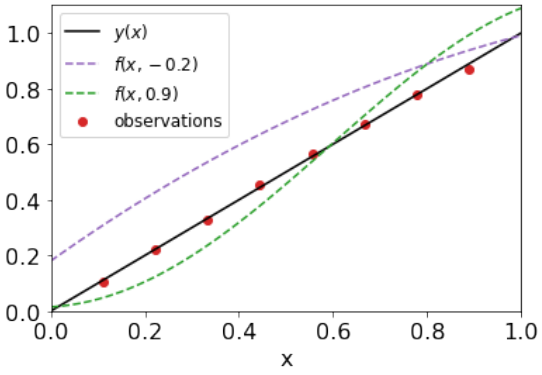


Figure 4.1: Observations, true process and predictions of the computer model with two different values of the parameter.

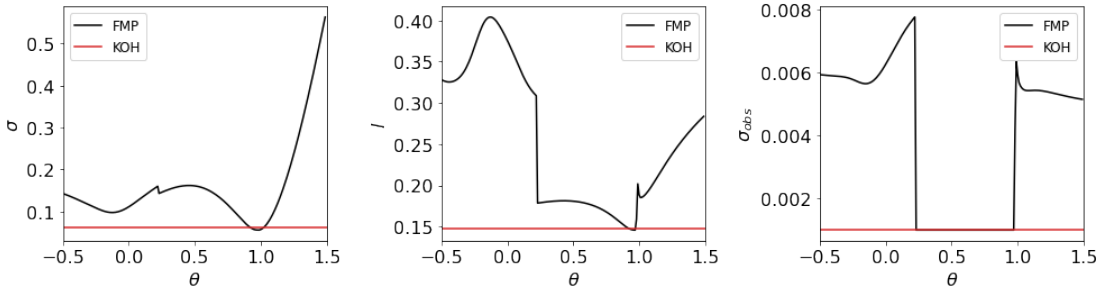


Figure 4.2: Estimators of hyperparameters for the FMP and KOH methods.

The hyperparameters to be learned about are σ, l and σ_{mes}. We take uniform priors for the model parameter, and uniform also for the two standard deviations on the range $[10^{-3}, 1]$ (zero is not included to avoid singularity). An inverse-gamma prior is taken for the correlation length: $l \sim IG(\alpha = 5.5, \beta = 0.3)$.

To perform the KOH method, the hyperparameters $\hat{\psi}_{KOH}$ must be estimated according to eq. (2.21), and the integral is computed using 200 integration points in the interval of θ. For the FMP method, an optimisation is computed at each step of the MCMC, and for the Bayes method, the M-H algorithm runs on a 4-dimensional space. The chains are run on 10^6 steps, with a 10% burn phase, and a total of 10^4 samples are extracted. For the Bayes chain, the total number of steps is $3e6$. To evaluate the good mixing of the chains, we compute the τ criterion of section 4.1.4, and we find $\tau(KOH) = 2.81$, $\tau(FMP) = 10.8$, $\tau(Bayes) = 159$, which shows that the increased dimensionality of the Bayes chain leads to a worse mixing. Thus, it requires more steps to get the same quality of samples.

The parameter posteriors are plotted on figure 4.3 and the optimal hyperparameters on figure 4.2. The reference posterior shows two peaks at $\theta \approx -0.2$ and $\theta \approx 0.9$, corresponding to two model predictions. In the KOH method, only the right peak is represented, and in the FMP method, both peaks are found, with misevaluated probability masses. In figure 4.2 it is shown that both values of θ correspond to different values of

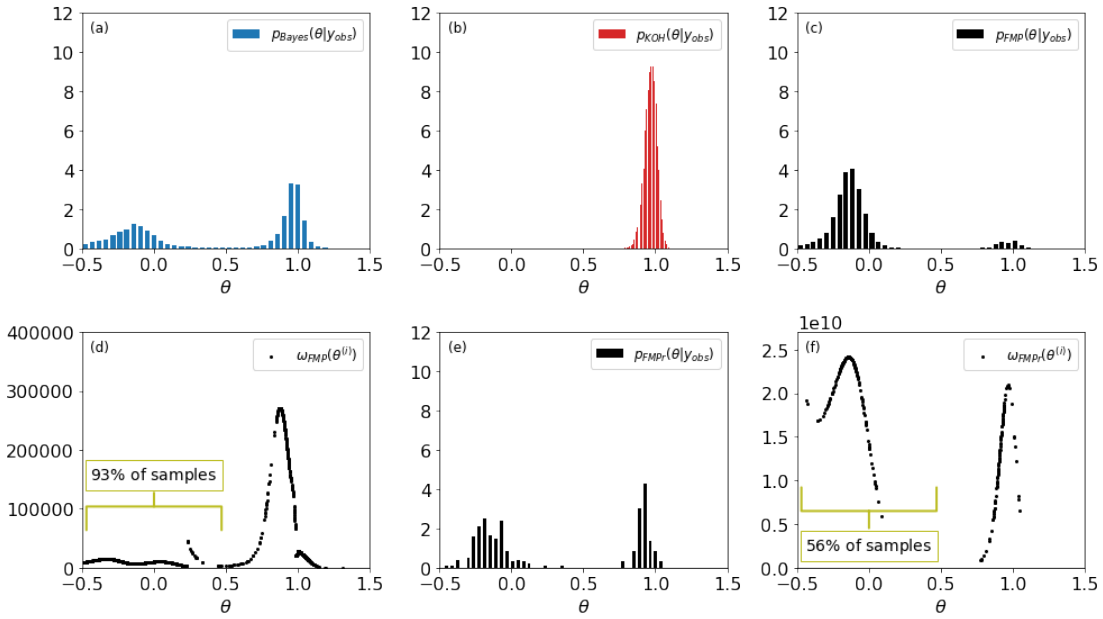


Figure 4.3: Posterior distributions of the model parameter with the Bayes method (a), KOH (b), FMP (c). (d): resampling weights calculated on the FMP sample. (e): FMPr sample obtained after resampling the FMP sample. (f): resampling weights calculated on the FMPr sample.

hyperparameters, and since in the KOH method, only one value can be estimated, only one peak is found. Note the effect of the prior of l, which places most of the probability mass on low values. Thus it is the reason why the right peak has such a high probability. The experimental error in both methods is slightly underestimated as some of it is explained by model error.

The resampling step of section 4.1.3 can be applied to improve the sample of the FMP posterior. The resampling weights are first computed for each point of the sample, and they are shown in figure 4.3 (top left). This plot reveals that the points above the value 0.5 have high resampling values, so it is a way to recognize that this region is undervalued. Next, we extract a sample of size 250 from the original sample of size 10^4 by sampling without replacement. Figure 4.3 shows possible results where five different samplings have been realized. Points on the right peak have been selected in the great majority, which gives a better balance between the intensity of the two peaks in the FMP method, at the price of reducing the sample size by a factor 40.

4.2 Surrogate modelling

In this thesis, surrogate modelling will be used to reduce the cost of expensive operations. It is applied to the computer code 4.2.1 and also to the FMP approximation 4.2.2.

4.2.1 Surrogate modelling of the computer code with dimensionality reduction

Until now, we have supposed that the output of the computer model was a scalar quantity, which could depend on a space variable. It is often the case that one run of the computer model produces a curve, of which we have a few observations. The model will provide a discretized version of the curve over a grid in most occurrences. Practically, one can see the model's output to be a vector quantity with the same size as the grid. If the output comprises multiple quantities (say a bubble radius and a temperature field), we can still see it as a vector with increased dimension. This section will discuss a technique to build a surrogate model of the code when the dimension of the output is high.

A natural approach is to interpolate the model's predictions on the observation points. The output dimension is then changed from the grid points to the observation points, which is a significant reduction in general. The issue is that the number of observation points can still be high, and also, one is often interested in the whole output of the model, not just at the observation points, which wouldn't be possible to represent with a surrogate model built just on those points.

To build a surrogate on the ensemble of grid points, we propose to use PCA on the output of the computer model and to select only a few modes that will carry the primary information of the data. There are multiple alternatives to PCA for emulating codes with high-dimensional output. One possibility is to find other bases of independent functions to build GPs on, which is referred to as the Linear Model of Coregionalization: see [Fricker et al., 2011] or [Bayarri et al., 2007] for wavelet bases. A popular field of application is for computer models with a time series output, as is discussed in [Conti and O'Hagan, 2010]. Another possibility is to treat the output dimensions as inputs of the GP and use elaborate high-dimensional, matrix-valued kernels to emulate the code, such as Separable Kernels [Micchelli and Pontil, 2004]. The central question thus becomes to find the best choice of kernel for a specific application. For a thorough review of emulation of vector-valued function using kernel methods, we refer to [Alvarez et al., 2012].

The full output of the code will be reconstructed as a linear combination of its principal modes, where the coefficients before the modes are scalar GPs. The procedure is detailed below.

Let d be the dimension of the output. Suppose that we have a nDoE consisting of s model runs, noted $D_1 = \{(\boldsymbol{\theta}_i, \mathbf{f}(\boldsymbol{\theta}_i))\}_{i<s}$. We note $\mathbf{Y}$ the $d \times s$ matrix where the i-th column is $\mathbf{f}(\boldsymbol{\theta}_i)$. Let $\bar{\mathbf{y}}$ be the mean of the $\{\mathbf{f}(\boldsymbol{\theta}_i)\}_{i<s}$, and let $\bar{\mathbf{Y}}$ be the matrix with all columns equal to $\bar{\mathbf{y}}$, so that $\widetilde{\mathbf{Y}} = \mathbf{Y} - \bar{\mathbf{Y}}$ is the centered data matrix. We note $\mathbf{C} = \widetilde{\mathbf{Y}}\widetilde{\mathbf{Y}}^T$ the data covariance matrix. The matrix $\mathbf{C}$ is symmetric definite positive, so we have:

$$\mathbf{C} = \mathbf{P}^T\mathbf{D}\mathbf{P}, \quad \text{with } \mathbf{P}^T\mathbf{P} = \mathbf{I}_d. \tag{4.18}$$

The columns of the matrix $\mathbf{P}$ are the eigenvectors $(\mathbf{e}_1, \dots, \mathbf{e}_d)$, arranged in decreasing order of eigenvalues. We select the first d_* eigenvectors, and we note $\mathbf{P}_*$ the $d \times d_*$ matrix with columns $(\mathbf{e}_1, \dots, \mathbf{e}_{d_*})$. The image space of the columns of $\mathbf{P}_*$ is referred to

as the *feature space*. Consider d_* independent scalar GPs, stored in the column vector $\mathbf{g}(\boldsymbol{\theta}) = (g_1(\boldsymbol{\theta}), \dots, g_{d_*}(\boldsymbol{\theta}))^T$. Thus the surrogate model for $\mathbf{f}$ reads:

$$\hat{\mathbf{f}}(\boldsymbol{\theta}) = \bar{\mathbf{y}} + \mathbf{P}_* \mathbf{g}(\boldsymbol{\theta}). \tag{4.19}$$

The training data for the GPs are the rows of the $d_* \times s$ matrix $\mathbf{P}_*^T \widetilde{\mathbf{Y}}$, which is the image of the original centered dataset in the feature space. Because the eigenvectors are orthogonal to each other, the GPs are chosen statistically independent [Wilkinson, 2010]. The training data can also be scaled using the eigenvalues or the data standard deviations, so the GPs can be built more efficiently.

The predictions at a given point $\boldsymbol{\theta}'$ follow a multivariate Gaussian distribution as a linear combination of Gaussians. We have :

$$\mathbb{E}\left[\hat{\mathbf{f}}(\boldsymbol{\theta}')|D_1\right] = \bar{\mathbf{y}} + \mathbf{P}_* \mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}')|D_1\right], \tag{4.20}$$

where $\mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}')|D_1\right] = (\mathbb{E}\left[g_1(\boldsymbol{\theta}')|D_1\right], \dots, \mathbb{E}\left[g_{d_*}(\boldsymbol{\theta}')|D_1\right])^T$ is the vector of the mean predictions of the scalar GPs. The total predictive variance writes:

$$\mathrm{Var}\left[\hat{\mathbf{f}}(\boldsymbol{\theta}')|D_1\right] = \mathbf{P}_* \mathrm{Var}\left[\mathbf{g}(\boldsymbol{\theta}')|D_1\right] {\mathbf{P}_*}^T, \tag{4.21}$$

where $\mathrm{Var}\left[\mathbf{g}(\boldsymbol{\theta}')|D_1\right] = diag(\mathrm{Var}\left[g_1(\boldsymbol{\theta}')|D_1\right], \dots, \mathrm{Var}\left[g_{d_*}(\boldsymbol{\theta}')|D_1\right])$ is the matrix of scalar variances, diagonal because the GP are independent. Note that the predictive variance is diagonal in the feature space but in the real space it is not, so the surrogate error is similar to a model error term, rather than an observation error term. The proposed surrogate model is stochastic, so its predictive variance is a natural representation of its uncertainty. We propose two ways to treat the surrogate uncertainty in the calibration framework:

- Consider that the surrogate uncertainty cannot be neglected and should be included in the calibration framework. This action is done by taking the mean prediction (4.20) as a deterministic surrogate (so it can be directly used in the calibration framework of section 3.1.2) and adding the predictive variance matrix to the matrices representing model error and observation error, in the likelihood function. This one represents the irreducible surrogate uncertainty that will logically impair the learning of model parameters.
- Estimate the surrogate uncertainty on a second nDoE (which can be done by splitting the first nDoE into a training dataset and a verification dataset), and if it is sufficiently low, do not include it into the calibration framework and simply use the mean prediction as a deterministic surrogate model.

The first approach was used for example in [Nagel et al., 2020; Roma et al., 2021]. The second approach requires more computer model evaluations, which can be prohibitive, but it also provides a natural way to choose the number of selected modes d_*. The calibration problem is also easier to solve (see [Higdon, Gattiker, et al., 2008]). We detail the second approach in the following paragraphs.

Suppose that we dispose of a verification nDoE of size s_{v}, noted $D_2 = \{(\boldsymbol{\theta}_{v,i}, \mathbf{f}(\boldsymbol{\theta}_{v,i}))\}_{i<s_{\mathrm{v}}}$ and the $d \times s_{\mathrm{v}}$ matrix $\mathbf{Y}_{\mathrm{v}}$ has for i-th column $\mathbf{f}(\boldsymbol{\theta}_{v,i})$. Let $\mathbf{Y}_{\mathrm{pred}}$ be the $d \times s_{\mathrm{v}}$ matrix with i-th column $\mathbb{E}\left[\hat{\mathbf{f}}(\boldsymbol{\theta}_{v,i})|D_1\right]$. We have $\mathbf{Y}_{\mathrm{pred}} = \bar{\mathbf{Y}} + \mathbf{P}_* \mathbf{G}$, where $\mathbf{G}$ is the $d^* \times s$ matrix with (i,j) coefficient equal to $g_i(\boldsymbol{\theta}_{v,j})$. We note $\mathbf{Y}_{\mathrm{proj}} = \bar{\mathbf{Y}} + \mathbf{P}_* \mathbf{P}_*^T (\mathbf{Y}_v - \bar{\mathbf{Y}})$. Note that $\mathbf{P}_* \mathbf{P}_*^T$ is the orthogonal projection onto the image space of $\mathbf{P}_*$, because $\mathbf{P}_*^T \mathbf{P}_* = \mathbf{I}_{d_*}$. Let $< \cdot, \cdot >$ be the usual scalar product in the space of matrices s.t. $< \mathbf{A}, \mathbf{B} >= Tr(\mathbf{A}^T \mathbf{B})$ and $|| \cdot ||$ the associated norm. Applying the Pythagorean theorem in the context of orthogonal projections:

$$\begin{aligned} ||\mathbf{Y}_{\mathrm{v}} - \mathbf{Y}_{\mathrm{pred}}||^2 &= ||(\mathbf{Y}_{\mathrm{v}} - \bar{\mathbf{Y}}) - (\mathbf{Y}_{\mathrm{pred}} - \bar{\mathbf{Y}})||^2 \\ &= ||(\mathbf{Y}_{\mathrm{v}} - \bar{\mathbf{Y}}) - \mathbf{P}_* \mathbf{P}_*^T (\mathbf{Y}_{\mathrm{v}} - \bar{\mathbf{Y}})||^2 + ||\mathbf{P}_* \mathbf{P}_*^T (\mathbf{Y}_{\mathrm{v}} - \bar{\mathbf{Y}}) - (\mathbf{Y}_{\mathrm{pred}} - \bar{\mathbf{Y}})||^2 \\ &= ||\mathbf{Y}_{\mathrm{v}} - \mathbf{Y}_{\mathrm{proj}}||^2 + ||\mathbf{Y}_{\mathrm{pred}} - \mathbf{Y}_{\mathrm{proj}}||^2, \end{aligned} \tag{4.22}$$

where we have used that $\mathbf{Y}_{\mathrm{pred}} - \bar{\mathbf{Y}}$ belongs to the image space of $\mathbf{P}_*$. Equation (4.22) reveals that the prediction error is the sum of two terms: the first one corresponds to the projection error, which is due to the projection on the feature space. The second term is the learning error, which is the residual error of using scalar Gaussian Processes with only a finite number of points.

Writing the uncertainty in this form gives us a rationale for choosing the number of selected modes d_*: as it increases, naturally the projection error decreases and the learning error increases (since more scalar GPs are added). Note that the cost of evaluating the surrogate model also increases with the number of scalar GP. As a general indicator, the optimal number of modes is taken when the uncertainty is half-split between the two sources. Another check is that the relative L2 error on the validation nDoE, given by:

$$\mathrm{e}_{\mathrm{L2}} = \frac{||\mathbf{Y}_{\mathrm{v}} - \mathbf{Y}_{\mathrm{pred}}||^2}{||\mathbf{Y}_{\mathrm{v}}||^2}, \tag{4.23}$$

should be of the order of a few percents to consider the surrogate model as an accurate representation of the computer code.

Another way to choose the number of modes is to compute the ratio of preserved energy, that is equal to $\sum_{i<d_*} \lambda_i / \sum_{i<d} \lambda_i$, where $\{\lambda_i\}_{i<d}$ are the ordered eigenvalues of $\mathbf{C}$. This ratio is understood as the preserved variance of the data and generally should not be lower than 95%.

4.2.2 Surrogate modelling in the FMP method: likelihood, or optimal hyparameters ?

Running a MCMC algorithm to get samples of the FMP approximation posterior requires a resolution of the optimisation problem at each step. Based on surrogate modelling, we propose two approaches to reduce the associated numerical cost.

- One approach is to build a surrogate model for each optimal hyperparameter. It requires one scalar GP per hyperparameter, and the input space of the GPs is the parameter space. We call each GPs dedicated to this use a hGP. In each step of the MCMC algorithm, the optimisation problem is replaced with the evaluation of these surrogates. Note that, at each step, the covariance matrix in the likelihood function still needs to be inverted.
- Another option is to build a direct surrogate model of the log-likelihood function (which we choose over the likelihood function for positivity constraints). We call this the "direct surrogate" technique, which requires one scalar GP with input space the parameter space. Now each step of the MCMC requires only a surrogate evaluation.

Note that it is relevant to build a surrogate model for the log-likelihood function because it has been proved continuous in section 3.2.3. The optimal hyperparameters have been proved to be upper-hemicontinuous which can create discontinuities as seen in section 4.1.5.

At first glance, the direct surrogate seems most advantageous because the matrix inversion is no longer required. Yet the question relevant to numerical cost is instead: How many training points are needed for each surrogate to have an acceptable precision? Suppose that the log-likelihood slowly varies with respect to parameters, then it is cheaper to build a surrogate on the function directly. On the contrary case of strong variations, maybe it would be easier to focus on the mapping $\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta})$ and use the exact likelihood function. Both approaches will be compared on an example in the next section.

Note that by choosing to use hGPs, we can obtain additional results that are not accessible directly under the second approach. For example, the second approach does not provide posterior distributions for the hyperparameters, which can give an insight into the nature of the model error (typically, there is a direct interpretation of variances or correlation lengths). Besides, if one wants to represent posterior predictions of the model with corrections, hyperparameters are necessary to compute posterior means of $\mathbf{z}_{\boldsymbol{\theta}}$. Consequently, obtaining these quantities under the second approach will require solving the optimisation problem many more times.

The hGPs can also be used not directly as optimal hyperparameters but as starting points for the optimisations. Thus, the number of optimisations would be the same as the full FMP method, but they would be shorter as they would get a good starting point. Even local optimisations could be fine if the hGPs are accurate. This feature adds to the numerical cost of the second approach, bringing it closer to full FMP, but solves the issue of some surrogate evaluations that might not be precise enough. This mix of methods is used for the calibration in chapter 5.

Building the surrogates for both methods also requires designing a nDoE. A simple option would be a one-shot design that fills well the prior space of model parameters. A preferred choice would be to locate training points in the regions where the posterior probability is high, but these regions are not known a priori. In section 4.3 we propose

an adaptative design of experiments which the goal of sampling these very regions.

4.2.3 Numerical example revisited

Let us compare the surrogate construction techniques to increase the performance of the FMP calibration on the previous numerical example of section 4.1.5. The gain in computation time will be examined, along with the quality of the surrogates.

Surrogate strategy

Let $D = \{\theta_i\}_{i<n_d}$ be a nDoE with size n_d. On each point, the computer model is evaluated and the optimal hyperparameters $\{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\theta_i)\}_{i<n_d}$ are computed. We consider two methods of surrogate construction: in the first one, at each point of the nDoE the values of the logprobability $\log \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\theta, \widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\theta))$ are computed. These values are used to build a surrogate model of the log-likelihood function that we note $\log \bar{\mathrm{p}}(\theta)$. The other approach is to use the data from the optimal hyperparameters to build three hGPs $(\widetilde{\sigma}(\theta), \widetilde{\sigma}_{obs}(\theta), \widetilde{l}(\theta))$. This results in a second surrogate model of the log-likelihood function given by $\log \widetilde{\mathrm{p}}(\theta) = \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\theta, \widetilde{\sigma}(\theta), \widetilde{\sigma}_{\mathrm{mes}}(\theta), \widetilde{l}(\theta))$.

Note that these surrogates aim to approach the FMP posterior and not the Bayes posterior. Previous notations have been shown how to construct a surrogate for the log-likelihood, which can be easily extended to the posterior by multiplication with prior parameters.

These surrogates can be readily used in the calibration framework to provide posterior samples and model predictions. We compare the runtime of each, along with the full FMP method. The quality of the approximations is evaluated using a posterior estimate. For the likelihood function, we compute the posterior L2 error with the following quantity:

$$\mathrm{fit}(\log \widetilde{\mathrm{p}}) = \frac{\sum_{\theta_i \in S} (\log \widetilde{\mathrm{p}}(\theta_i) - \log \mathrm{p}_{\mathrm{FMP}}(\theta_i))^2}{\sum_{\theta_i \in S} (\log \mathrm{p}_{\mathrm{FMP}}(\theta_i))^2}, \tag{4.24}$$

where S is the sample obtained from the full FMP calibration. The fit quality for the hyperparameters is estimated as:

$$\mathrm{fit}(\widetilde{\psi}) = \frac{\sum_{\theta_i \in S} (\widetilde{\psi}(\theta_i) - \hat{\psi}_{FMP}(\theta_i))^2}{\sum_{\theta_i \in S} (\hat{\psi}_{\mathrm{FMP}}(\theta_i))^2}. \tag{4.25}$$

Numerical example

We go back to the example of section 4.1.5 and apply the surrogate construction techniques to accelerate the FMP calibration. The nDoE is calculated following a Halton sequence with size varying from $n_d = 50$ to $n_d = 350$. The hGPs are constructed with mean zero and with Matern 3/2 kernels. The corresponding hyperparameters are estimated using Maximum Likelihood.

Method	Steps			Total
	Surrogate construction	MCMC	Predictions	
with hGP, $n_d = 50$	50	8	27	85
with hGP, $n_d = 350$	352	31	32	415
Surrogate of p_{FMP}, $n_d = 50$	50	2	27	79
Surrogate of p_{FMP}, $n_d = 350$	350	9	32	391
Full FMP	0	1196	28	1224

Table 4.1: Detail of the computation time (in seconds) for each calibration method.

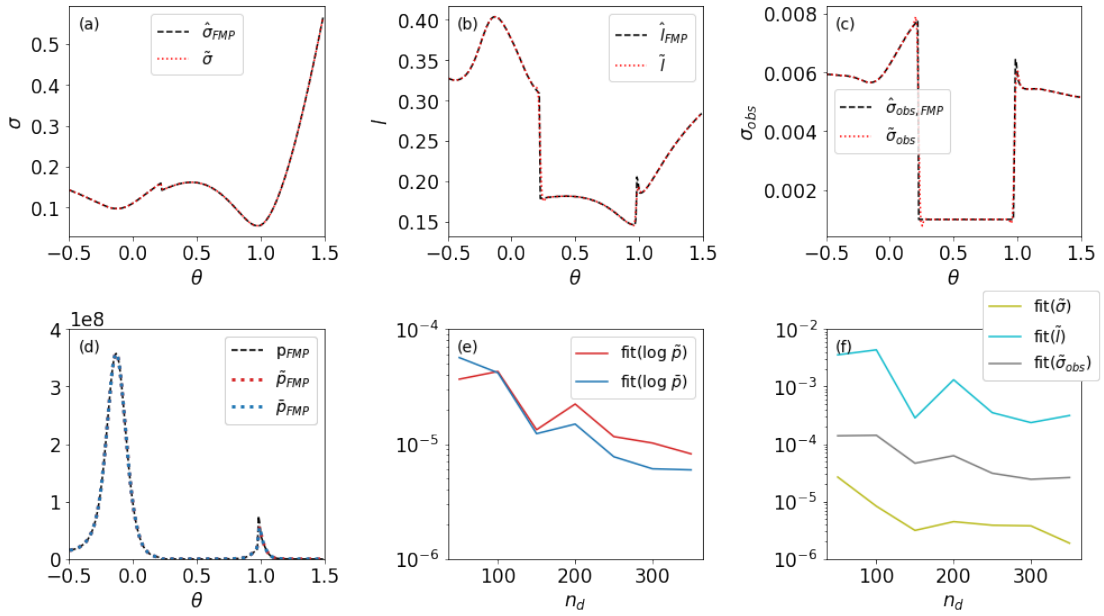


Figure 4.4: (a), (b), (c): hGPs with $n_d = 350$. (d): Surrogates of the likelihood function with $n_d = 350$. (e): Error criteria on the log-likelihood function. (f): Error criteria on the hyperparameters.

The computation time is detailed in table 4.1, where it can be seen that the construction of the surrogates greatly improves the total calibration time. The MCMC is noticeably faster when using a direct surrogate. Note that for the resolution of the optimisation problem, we can set a maximum research time that is noted t_{opt}. For the MCMC of the full FMP we chose $t_{\text{opt}} = 10^{-4}s$ to keep a reasonable computation time. The optimisations for the predictions are also set with $t_{\text{opt}} = 10^{-4}s$ because the size of the posterior sample is 10^4. For the initial optimisations that are used to build the surrogates, we set $t_{\text{opt}} = 1s$ to get better precision. Note that using a longer research time for the optimisation allows better optimal points, but it does not influence the maximal value attained. All MCMCs are well-mixed according to the mixing criterion.

In figure 4.4 surrogates are plotted along with target functions, and error criteria are also shown as a function of the size of the nDoE. The surrogates accurately represent the target functions, so the samples from all three posterior distributions are similar, and

the calibration techniques lead to the same results.

It can be argued that the log-likelihood function behaves more nicely than the optimal hyperparameters, so that building a direct surrogate would be more accessible. Besides, building a direct surrogate is a more targeted approach, leading to a more accurate approximation, as confirmed by the fit criteria. Despite the discontinuities, the surrogates for hyperparameters still provide a good representation of the log-likelihood function.

In this example, an optimisation time of $t = 10^4 s$ is sufficient to get a good precision on the FMP method. In cases with a more significant number of observations or a higher dimension of hyperparameters, this time would need to be increased, underlining the necessity of using surrogate models. This aspect would also increase the time required for predictions when using a direct surrogate, for the reasons mentioned in section 4.2.2. These conclusions lead us to use surrogates for optimal hyperparameters in the rest of this work.

4.3 Adaptive design of computer experiments

We present an adaptive algorithm for choosing the best locations to build the hGPs. The rationale of this algorithm is to place training points in regions where the posterior probability is high. Such a strategy has already been employed for the construction of the computer model surrogate, for example, in [Conrad et al., 2016] using gradually refined local approximations of the computer code, or in [Damblin and Gaillard, 2019] where an Expected Improvement maximisation problem is solved on the parameter space to place training points. Here we adopt the view of [Lucor and Maitre, 2018], where they aim to use posterior samples as training points for the surrogates. It fills the objective of high posterior probability, but also it can place points in regions that are incorrectly assumed to have low posterior probability, thus allowing exploration of the parameter space. In this work, we suppose that the surrogate for the computer model has been built in a previous step and can be used freely. For another example of adaptive algorithms applied to calibration problems, we refer to [Pratola et al., 2013] who proposed to build a surrogate model of a discrepancy criterion between the predictions and the observations, although without consideration of model error.

In section 4.3.1, we explain the rationale for choosing posterior plausible points by showing that the quality of approximation of the posterior is bounded by the quality of the hGPs. In section 4.3.2, we present the adaptive design algorithm, and in section 4.3.3 this approach is compared to traditional fixed design methods on a numerical example with real data, borrowed from chapter 6.

4.3.1 Why choose points that are plausible a posteriori?

Here we give proof that increasing the precision of the surrogates in the region where the posterior probability of parameters is high will lead to a better approximation of the FMP posterior. The fit between two distributions is measured as previously by the KL

divergence. We follow the methodology of [Marzouk and Xiu, 2009], who proved that the KL divergence was upper bounded by the surrogate model error induced by the prior measure, using a Polynomial Chaos surrogate for the computer model.

For clarity in this section we take the following notations: $L_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) = \log \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}))$ is the FMP log-likelihood function. We note $\widetilde{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta})$ the hGPs and $\widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) = \log \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \widetilde{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}))$ the log-likelihood function in which the hGPs are plugged in. The prior distribution for the model parameters is $\text{p}(\boldsymbol{\theta})$, the marginal likelihoods are noted $\gamma = \int_{\boldsymbol{\theta}} \exp(L_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\, \text{p}(\boldsymbol{\theta})\ \text{d}\boldsymbol{\theta}$ and $\widetilde{\gamma} = \int_{\boldsymbol{\theta}} \exp(\widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\, \text{p}(\boldsymbol{\theta})\ \text{d}\boldsymbol{\theta}$. According to Bayes' Theorem we have:

$$\text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = \text{p}(\boldsymbol{\theta})\frac{\exp(L_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))}{\gamma}, \quad \widetilde{\text{p}}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) = \text{p}(\boldsymbol{\theta})\frac{\exp(\widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))}{\widetilde{gamma}}.$$

The proof relies on two assumptions:

- There is little prior information on hyperparameters, so that $\widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}) \approx \arg\max_{\boldsymbol{\psi}} \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi})$.
- The hyperparameter derivatives of the log-likelihood are bounded: $\forall(\boldsymbol{\psi}, \boldsymbol{\psi}') \in \Psi^2, \exists \Lambda > 0, |\log \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi}) - \log \text{p}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi}')| \leq \Lambda ||\boldsymbol{\psi} - \boldsymbol{\psi}'||_{L2}$. Note that this is verified by the Matern family for instance when the hyperparameter search space Ψ is compact, because the derivatives are continuous (see section 3.2.2).

Since the behaviour of the likelihood function is highly dependent on the covariance family and the choice of varying hyperparameters, the second assumption, which concerns the regularity of this function, is required. We note, for a function $f : \boldsymbol{\theta} \mapsto f(\boldsymbol{\theta})$ the posterior L2 norm: $||f||^2_{L2,\text{p}_{\text{FMP}}(\boldsymbol{\theta})} = \int_{\boldsymbol{\theta}} f^2(\boldsymbol{\theta}) \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})\ \text{d}\boldsymbol{\theta}$. The quality of the approximation is written as the KL divergence:

$$\begin{aligned} DKL(\text{p}_{\text{FMP}}||\widetilde{\text{p}}_{\text{FMP}}) &= \int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) \log \frac{\text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})}{\widetilde{\text{p}}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})}\ \text{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) \log \frac{\exp(L_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\widetilde{\gamma}}{\exp(\widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\gamma}\ \text{d}\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}})(L(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) - \widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\ \text{d}\boldsymbol{\theta} + \log \frac{\widetilde{\gamma}}{\gamma}. \end{aligned} \tag{4.26}$$

According to the first assumption, we have $L_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) \geq \widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta})$ for $\boldsymbol{\theta} \in \Theta$, thus $\gamma \geq \widetilde{\gamma}$. Consequently, the first term in the right-hand side of eq. (4.26) is positive and the second is negative. Thus, the KL divergence can be controlled by an upper bound on the positive term:

$$\begin{aligned} D_{\text{KL}}(\text{p}_{\text{FMP}}||\widetilde{\text{p}}_{\text{FMP}}) &\leq |\int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) * (L(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) - \widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}))\ \text{d}\boldsymbol{\theta}| \\ &\leq \int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) * |L(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta}) - \widetilde{L}_{\text{FMP}}(\mathbf{y}_{\text{obs}}|\boldsymbol{\theta})|\ \text{d}\boldsymbol{\theta} \\ &\leq \Lambda \int_{\boldsymbol{\theta}} \text{p}_{\text{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\text{obs}}) * ||\widehat{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta}) - \widetilde{\boldsymbol{\psi}}_{\text{FMP}}(\boldsymbol{\theta})||_{L2}\ \text{d}\boldsymbol{\theta}, \end{aligned} \tag{4.27}$$

where we have used the assumption of bounded derivatives.

Equation (4.27) gives the rationale of the algorithm, which is to make the upper bound as low as possible. From this, we derive two objectives:

- **Objective 1 (exploitation):** place training points in regions of high probability of $\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ to improve the precision of the surrogates $\widetilde{\boldsymbol{\psi}}_{\mathrm{FMP}}$ in these regions,
- **Objective 2 (exploration):** place training points in regions where $\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta}|\mathbf{y}_{\mathrm{obs}})$ is not precisely known, in order to find all regions of high probability.

4.3.2 The sampling algorithm

The proposed algorithm is a sequential procedure in which we alternate between three phases:

1. The construction of the surrogate $\widetilde{\boldsymbol{\psi}}_{\mathrm{FMP}}$ which leads to an approximation $\widetilde{\mathrm{p}}_{\mathrm{FMP}}$,
2. The sampling of $\widetilde{\mathrm{p}}_{\mathrm{FMP}}$ with a classical MCMC algorithm and the extraction of a set of candidate training points,
3. The choice of new training points in the set of candidate points, operated by resampling.

The goal is to get an increasingly refined approximation of $\mathrm{p}_{\mathrm{FMP}}$.

We note $\widetilde{\boldsymbol{\psi}}^{(n)}_{\mathrm{FMP}}(\boldsymbol{\theta})$ the surrogate for hyperparameters at step n, $\widetilde{L}^{(n)}_{\mathrm{FMP}}(\boldsymbol{\theta}) = \log \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \widetilde{\boldsymbol{\psi}}^{(n)}_{\mathrm{FMP}}(\boldsymbol{\theta}))$ the resulting approximation of the log-likelihood function. The surrogate is built on the set of training points of the step $n-1$, which is noted $T^{(n-1)} = \{\boldsymbol{\theta}^t_i\}_{i \leq t_{n-1}}$. The sampling phase of step n consists in running a M-H algorithm to obtain a sample of the density proportional to $\mathrm{p}(\boldsymbol{\theta}) \exp(\widetilde{L}^{(n)}_{\mathrm{FMP}}(\boldsymbol{\theta}))$. This is the set of candidate points for the step n, which is noted $C^{(n)} = \{\boldsymbol{\theta}^c_i\}_{i \leq c_n}$.

Their predictive variance gives a good estimate for the prediction error of the surrogates. We use the following:

$$||\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}) - \widetilde{\boldsymbol{\psi}}^{(n)}_{\mathrm{FMP}}(\boldsymbol{\theta})||^2_{L2} \approx e^{(n)}_{\mathrm{pred}}(\boldsymbol{\theta}) := \sum_{i \leq h} \frac{\mathrm{Var}\left[\widetilde{\psi}^{(n)}_{i,\mathrm{FMP}}(\boldsymbol{\theta})|T^{(n-1)}\right]}{\sigma_i^2}, \tag{4.28}$$

where $\widetilde{\psi}^{(n)}_{i,\mathrm{FMP}}$ is the hGP for the i-th hyperparameter at step n, and σ_i is the standard deviation of the i-th hyperparameter calculated on the training set.

The resampling procedure, presented in another context in section 4.1.3, consists of a weighted random sampling, without replacement, among the candidate points. The new sample is noted $R^{(n)}$, with size $r_n \ll c_n$. We adopt the rule of thumb $r_n/c_n \approx 1/20$, as the new sample size must be significantly reduced to see the benefits of the resampling procedure. The resampling weights are given by:

$$\omega_i = \sqrt{e^{(n)}_{\mathrm{pred}}(\boldsymbol{\theta})}. \tag{4.29}$$

The effect of these weights is to favor regions where the predictive variance is high, following the exploration principle. Note that, in regions where the posterior probability is estimated to be high, there will be a significant number of candidate points; thus, the chance of sampling this region is exacerbated in accordance with the exploitation objective. The complete procedure is presented in algorithm 2.

Algorithm 2 Selection of training points for the hGPs

1: Initialization: if $n = 0$, pick an initial training set T^{init} as a space-filling design (LHS or QMC).
2: **for** $n < n_{\max}$ **do**
3: Update the surrogate $\widetilde{\psi}^{(n)}_{i,\text{FMP}}$ on the current training set $T^{(n-1)}$.
4: Obtain a sample $C^{(n)}$ of the density $\mathrm{p}(\boldsymbol{\theta})\exp(\widetilde{L}^{(n)}_{\text{FMP}}(\boldsymbol{\theta}))$ by running a MCMC algorithm.
5: Compute the weights $\{\omega_i\}$ for each point of $C^{(n)}$. Set $R^{(n)} = \emptyset$.
6: **for** $i \leq r_n$ **do**
7: Perform a weighted draw of the point $\boldsymbol{\theta}^c_j$ from $C^{(n)}$, add $\boldsymbol{\theta}^c_j$ to $R^{(n)}$, and set $\omega_j = 0$.
8: **end for**
9: Evaluate $\widehat{\boldsymbol{\psi}}_{\text{FMP}}$ on each point of $R^{(n)}$ and set $T^{(n)} = T^{(n-1)} \cup R^{(n)}$.
10: **end for**

Note that the user must specify the number of points added to the training set at each step. At the beginning of the algorithm, many points must be added to favor exploration of the high probability zones. Later, when the surrogates are believed to be precise enough, this number can be reduced. Also, keep in mind that for Gaussian Processes, the training set must be of limited size (say thousands) because of the evaluation cost. These limits can be overcome with sparse GP techniques [Bauer et al., 2017], or by gradually removing points from the training set [Lucor and Maitre, 2018], but it is not envisaged here. In our DEBORA application, it is checked that a good precision is achieved with a training set of size 500, which we obtain by running $n_{\max} = 2$ MCMCs for computational efficiency.

4.3.3 Precision study of the adaptive algorithm

In this section, we examine the performance of the proposed adaptive algorithm compared to offline space-filling techniques such as QMC and LHS. We will show that, at the additional cost of sampling a few extra MCMCs and running the resampling algorithm, we can achieve a better precision of the surrogates for the optimal hyperparameters.

This exercise is done on real data, borrowed in advance from chapter 6. The observations are constituted of experimental measurements of both void fraction and bubble diameter, obtained from the A6 case of the DEBORA experiments. In this section, we focus only on the sampling algorithms so that discussions about the data and the calibration assumptions are deferred to section 6.5. The only requirement for this study is that we can compute the optimal hyperparameters $\widehat{\boldsymbol{\psi}}_{\text{FMP}}$ for a given $\boldsymbol{\theta}$, as well as the

likelihood function $\mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta},\boldsymbol{\psi})$ for any value of $(\boldsymbol{\theta},\boldsymbol{\psi})$. In this example, $\dim(\boldsymbol{\psi}) = 6$ and $\dim(\boldsymbol{\theta}) = 5$.

We also use a reference sample of parameters $\Theta_s = \{\boldsymbol{\theta}_i\}_{i\leq s}$, with size $s = 3000$, that was obtained from the true FMP distribution $\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta})$ in this problem. The corresponding values of optimal hyperparameters $\{\widehat{\boldsymbol{\psi}}_{\mathrm{FMP}}(\boldsymbol{\theta}_i)\}_{i\leq s}$ are known. From this sample, it is possible to build a reference procedure for the construction of the hGPs: the training points would be progressively taken from the reference sample. Thus it is relevant to compare the rates of convergence of both methods.

From the reference data we can compute a quantitative criterion for the evaluation of a given surrogate model $\widetilde{\boldsymbol{\psi}}_{\mathrm{FMP}}$ as:

$$e_{i,\mathrm{L2}}(\widetilde{\boldsymbol{\psi}}_{\mathrm{FMP}}) = 100 \times \sqrt{\frac{\sum_{\boldsymbol{\theta}\in\Theta_s}(\widetilde{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta}) - \widehat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta}))^2}{\sum_{\boldsymbol{\theta}\in\Theta_s}(\widehat{\boldsymbol{\psi}}_{i,\mathrm{FMP}}(\boldsymbol{\theta}))^2}}. \tag{4.30}$$

The adaptive algorithm was tested with three values $r_n = 50, 100, 200$, which correspond to the number of points added in the training set at each iteration. The relative error on the surrogates is plotted in figure 4.5. This figure shows that the precision of the adaptive algorithm is on par with the reference solution for each hyperparameter, and it outperforms space-filling designs. The choice of r_n bears no influence on the precision, so the highest value will be preferred in the applications because less MCMCs will be required to run. The error criterion is an estimator for the posterior-averaged squared error on the hyperparameter value, which in the case of $\boldsymbol{\psi}_4$ is around 2% for the best-trained models. This order of magnitude for the error does not impact the likelihood values, as we will see on the marginals plots.

To evaluate how the precision of the surrogates impact the precision of the overall likelihood function, the successive approximations of the likelihood function were plotted on the middle row of figure 4.6, by performing a KDE on the candidate set of parameters obtained at each step. The top plot in the figure corresponds to the true FMP likelihood function, obtained with a KDE on the reference sample. On Step 0, the likelihood approximation is obtained from a space-filling nDoE of size 50, and at each step, we add $r_n = 200$ points to the training set. A scatterplot on the bottom row where each blue marker represents a point in the candidate set and a red marker represents a selected point. The markers are placed depending on the value of the likelihood function and the value of their resampling weight. This plot allows visualizing that most of the selected points are in the high probability region while permitting the choice of low-probability points with high predictive uncertainty. As the likelihood function is better approximated, the samples in the candidate set become similar to the reference sample. The number of candidate points with a high uncertainty decreases sharply, and as a consequence, most of them are included in the training set. Eventually, all high-probability regions are sufficiently explored, and the approximation of the likelihood function becomes valid.

From this study, we conclude that the proposed adaptive algorithm for building

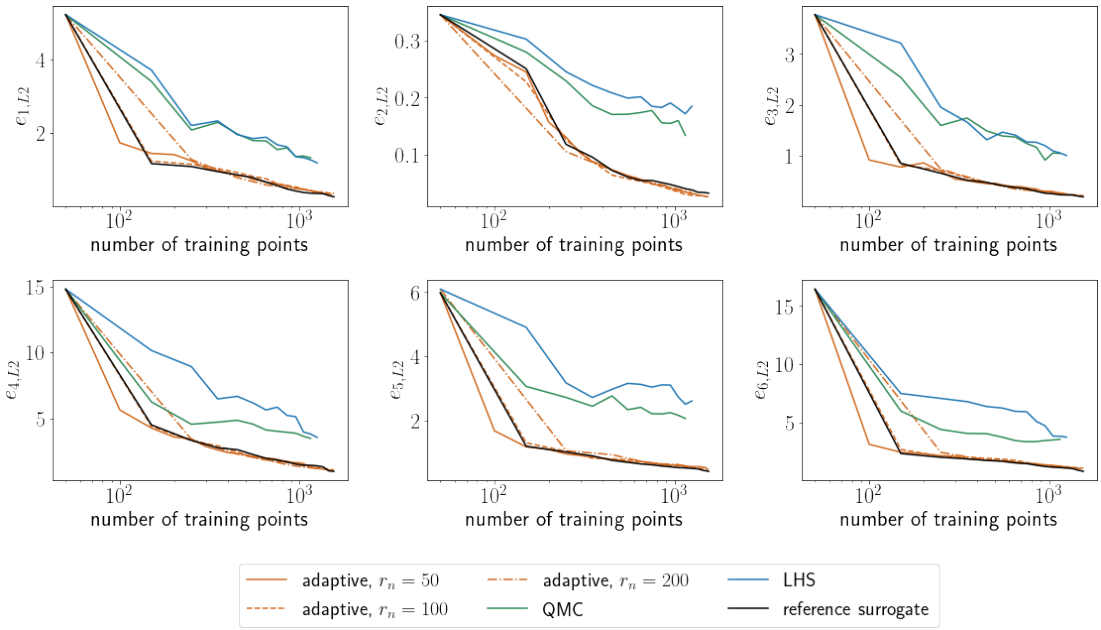


Figure 4.5: Posterior L2 error for the hyperparameter surrogates as a function of the number of training points, for different algorithms. Each plot corresponds to one hyperparameter.

hyperparameter surrogates outperforms offline techniques such as QMC or LHS designs. It is easy to implement, and the effective overhead is the cost of running a low number of supplementary MCMCs. Due to the choice of training points in high posterior regions, while allowing for exploration of the parameter space, it will enable to get a surrogate likelihood that is precise enough for calibration purposes.

4.4 Conclusion

In this chapter, we have introduced numerical techniques that can be used to perform efficient calibration of computer models. The use of MCMC techniques is prime in Bayesian numerics, so we have detailed how it can be applied in the context of KOH, FMP, and Bayes calibration, along with convergence diagnostics and sample extraction. Surrogate models are essential when the computer code is long to run, and we have presented a classical construction technique in conjunction with PCA when the output of the code is high. The cost of the FMP calibration being too high for practical applications due to the significant number of optimisations required, so we have compared two techniques based on surrogate models to correct this issue. Finally, a simple adaptive algorithm for picking training points that are plausible a posteriori has been proposed to reduce the number of optimisations required in a calibration problem. The precision study on real data shows that this will be sufficient for practical applications.

Adaptive algorithms for the choice of training points are an active area of research in UQ. Trying to use points in the posterior plausible region is a popular strategy. Using

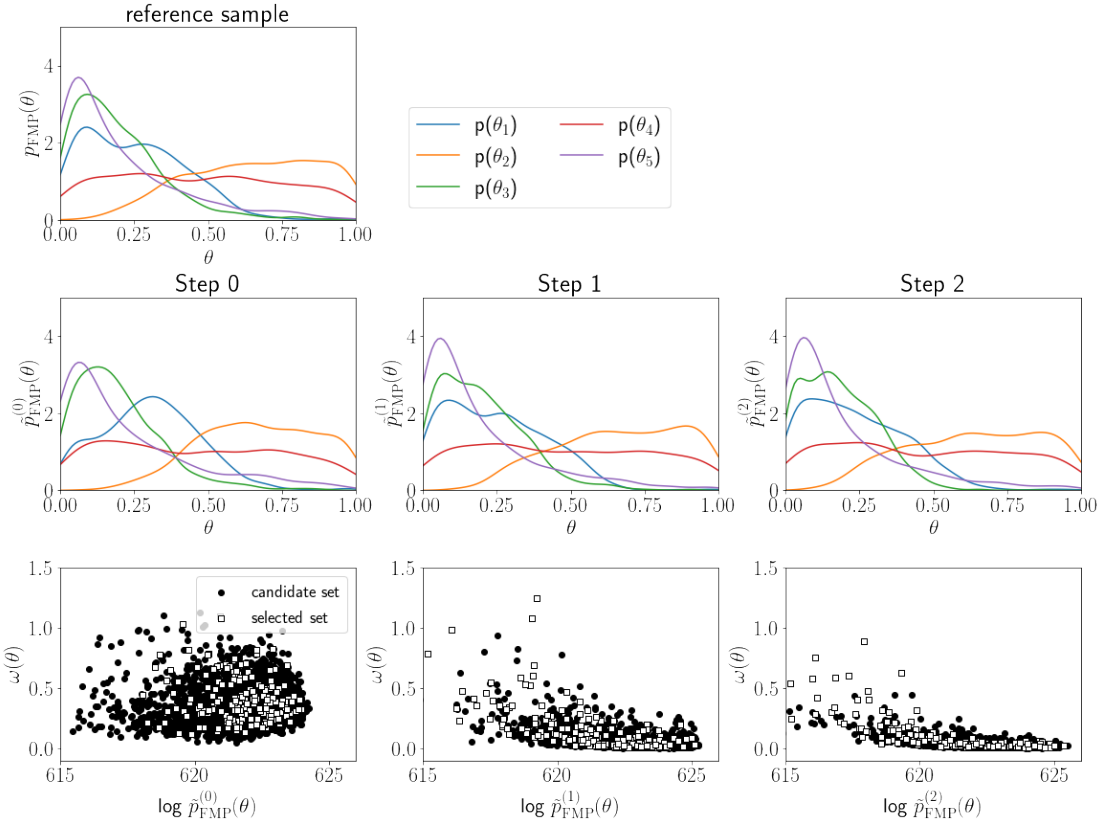


Figure 4.6: Plot of the marginals for each calibrated parameter, obtained by KDE. Top row: reference distribution $\mathrm{p}_{\mathrm{FMP}}(\boldsymbol{\theta})$. Middle row: successive distributions of the training set, with increasing number of training points for the surrogates. Bottom row: successive distributions of the candidate set, with increasing number of training points.

a resampling step in our proposed algorithm requires throwing away many samples that could be expensive to acquire, although, in our applications, we have not encountered this issue. Thus, an alternative way of choosing the training points could be to minimize the posterior-averaged prediction variance over the parameter space, serving both as exploration and exploitation, in the spirit of Expected Improvement. In [Damblin, Barbillon, et al., 2018], such techniques are employed for the construction of a surrogate model for the computer code directly, which also constitutes a natural extension of our algorithm, as the consideration of model error could improve the guided design. Finally, these adaptive designs can be employed for building surrogate models other than GPs, as applied in [Lucor and Maitre, 2018] to Polynomial Chaos.

Chapter 5

Application to a heat flux partitioning model

Contents

In this chapter, we consider the boiling of water in contact with a heated wall in various flow conditions. MITB studied in this chapter can be used to compute the components of the heat flux from the wall to the water as a function of the wall temperature. Our goal is to use data from a boiling experiment to improve the knowledge of empirically determined coefficients in the model and the contact angle between water and the wall, which is notoriously difficult to estimate in high temperatures and pressures. This calibration will also provide credible intervals (Bayesian equivalent to confidence intervals) on the model predictions, leading to a more accurate data assessment. In the first step, the calibration is done on each experiment separately to check the individual behaviours. Then, a global calibration including all experiments is performed to assess the uncertainty globally.

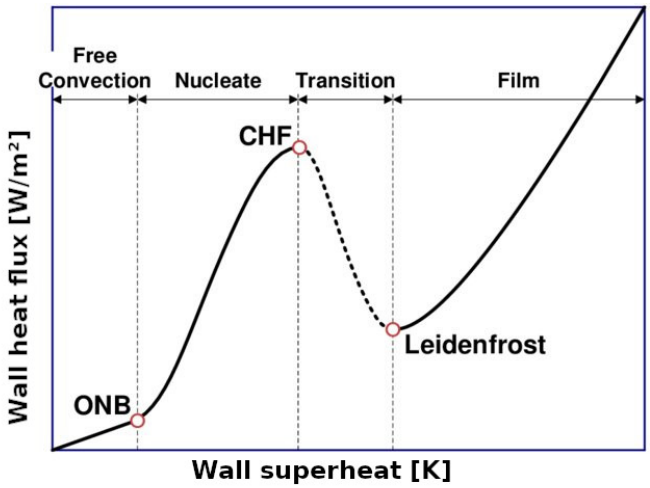


Figure 5.1: Nukiyama boiling curve, from [Liu et al., 2015].

5.1 Physical phenomenon: heat flux in wall boiling

Thermal engineering problems can be generally summed up by the question: "How can we cool down this given object and do it efficiently?". A simple solution for cooling, for example, a hot metal cube, is to drop it in a pool of cold water. The same principle is used in nuclear engineering: the heating element, which is the nuclear reactor, is cooled by the primary water circuit that surrounds it and transfers the heat to the secondary water circuit, which is then used for electricity production. The cooling efficiency of reactors is determined by parameters in the primary circuit such as the water temperature, its pressure, or the geometry of the contact. Another central question that distinguishes between two reactors families is determining whether or not the water should be boiling.

It turns out that boiling is a highly efficient way of transferring heat, as the transformation of liquid water into vapor requires a significant quantity of energy to be extracted from the reactor, which as a result, is significantly cooled. In contrast, if the water remains below the boiling point, the energy from the reactor will only serve to elevate the liquid temperature, which is not as efficient. This behaviour is well-known and represented on the two first sections of the boiling curve of figure 5.1, established by Nukiyama [Nukiyama, 1966], representing the quantity of heat transfer against the wall superheat, which is the temperature difference between the wall and the boiling temperature (also saturation temperature) of the liquid. Before boiling occurs, the heat transfer is fully convective. The first bubbles appear at the ONB and the heat transfer increases significantly in the nucleate boiling regime.

If the wall superheat increases even further, the CHF point is reached (see figure 5.1): the bubbles at the wall surface begin to form a thin film of vapor that surrounds the wall, which is the start of the film boiling regime. This phenomenon results in a jump to the Leidenfrost point on the figure, which is accompanied by a sudden drop in heat flux

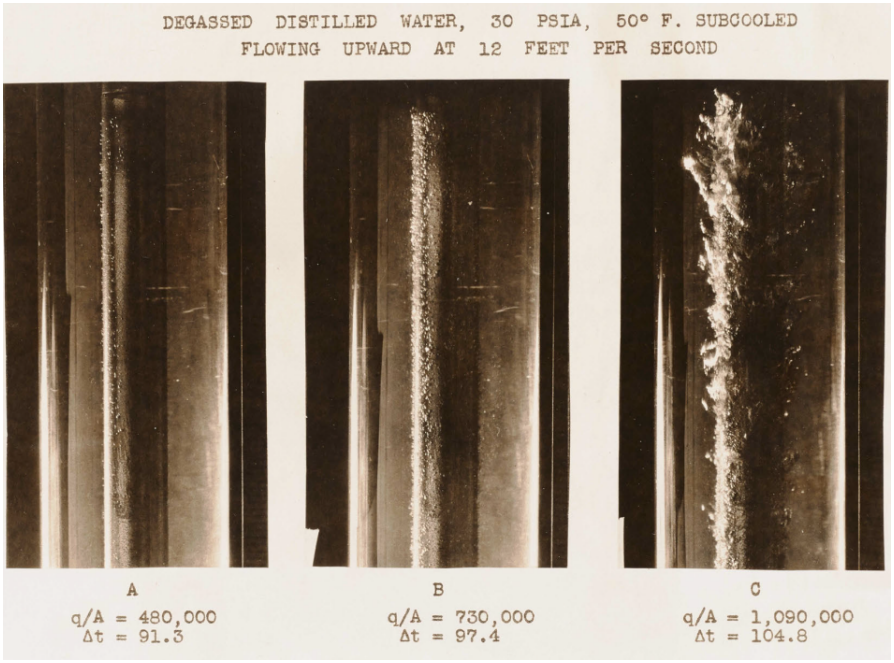


Figure 5.2: Photographs of the Kennel experiment, from [Kennel, 1949]. Left: ONB. Middle: nucleate boiling. Right: transition to film boiling (critical heat flux reached).

efficiency: since a gas conducts heat way slower than a liquid, it creates insulation of the reactor, which can lead to overheating, with potentially significant consequences. Thus, it is crucial in energetics applications to be able to compute the wall heat flux in the convection and nucleate regimes and to accurately predict the value of the CHF. Some photographs of the boiling experiment that will provide the data in this chapter are shown in figure 5.2, providing a visualisation of different water regimes boiling around a heated rod.

It is noted that in the nuclear field, the principles of cooling using boiling are applied in the design of BWRs, yet the most common type of reactor is the BWR, in which the cooling of the reactor remains in the convection regime, as the water remains liquid due to an imposed pressure of 155 bar. In PWRs, the cooling efficiency is traded for better safety (no risk of CHF in operating conditions) and a higher operating temperature which results in more energy production. Yet, the CHF is of uttermost importance in thermal-hydraulics of PWRs for two reasons: first, in operating conditions, there is boiling at locations other than the reactor, for example, in the secondary circuit. Secondly, some accidental scenarios lead to a loss of pressure in the primary circuit, lowering the saturation temperature of the water and potentially inducing boiling at the reactor. Thus, the nuclear context is a strong catalyzer for research works directed towards a better understanding of the mechanisms revolving around heat transfer in the boiling of liquids.

The repercussions of the boiling regime on the overall liquid flow, for example, concerning the quantity of gas produced or bubble interaction mechanisms, will be the

subject of the application in chapter 6.

5.2 Mechanistic model for heat flux prediction

The model that is considered in this chapter is the MITB of Kommajosyula [Kommajosyula, 2020], which is used in boiling studies to compute the heat flux from the wall to the liquid (wall heat flux ϕ) as a function of the wall superheat ΔT_{sup}, which is the difference between the wall temperature T_{w} and the saturation temperature T_{sat}. The wall heat flux also depends on the experimental conditions such as the bulk temperature T_{bulk}, the liquid pressure p, the flow rate Q and the hydraulic diameter D_{h}.

Running the MITB is typically inexpensive because it requires only the evaluation of a few numerical expressions. It was initially developed in MATLAB, using the package XSteam to obtain the thermodynamical properties of fluids in an extensive range of pressures and temperatures. For this study, the model was rewritten in C++. All necessary steps to reproduce the code can be found in Appendix B of [Kommajosyula, 2020]. Because of its semi-empiric nature, the model is extremely cheap to evaluate, and thus we do not need a surrogate model.

In section 5.2.1, we give a brief overview of the physical principles underlying the model. In section 5.2.2, we detail two experimental correlations used in the model that will be the target of our calibration study. Finally, in section 5.2.3 the Kennel experiment is presented along with the performance of the MITB on it.

5.2.1 Base physical principles

MITB is based on the model of Gilman [Gilman and Baglietto, 2017], which takes into account complex bubble-wall interactions such as the inception (as the bubble grows in a cavity), the microlayer evaporation (as the bubbles detach), the quenching (liquid filling the void left by bubbles sliding at the wall), and the reduction of the predicted number of nucleation sites due to their absorption by other bubbles. These mechanisms are represented in figure 5.3.

The model consists in assembling, in a physically grounded manner, quantities that are central to these mechanisms, for example:

- The bubble departure frequency f_{dep} which is the rate at which bubbles nucleate, grow, and depart the wall,
- The departure diameter D_{d} which is the bubble diameter at lift-off from the surface,
- The contact angle θ_{c} which is the static angle between the wall and the bubble membrane before lift-off.

These quantities are usually obtained from experimental correlations involving adimensional numbers, with a structure prescribed by the modeller and coefficients obtained by calibration against experimental data. Either a direct correlation is used, or they are predicted by assembling correlations for other quantities. A common correlation for f_{dep}

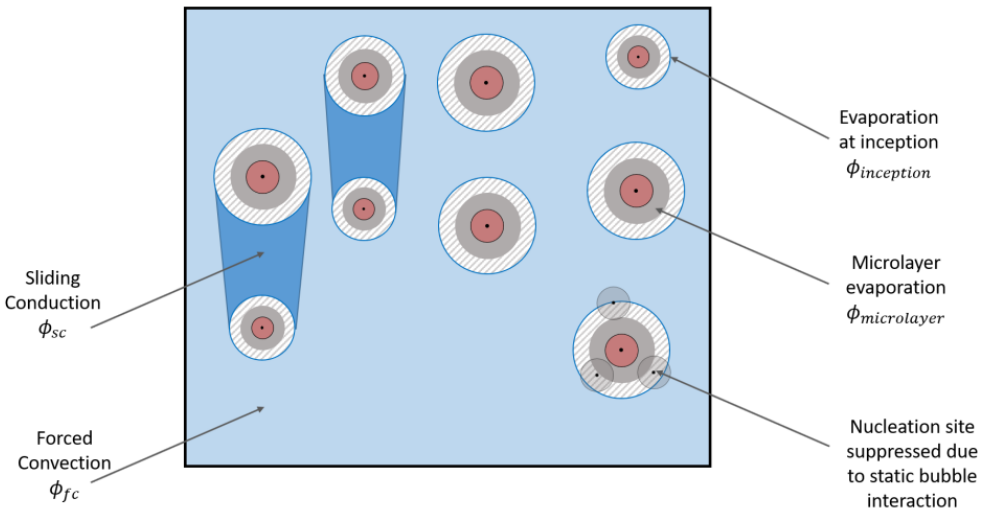


Figure 5.3: Illustration of the bubble-wall interactions and types of heat transfer considered in the MIT Boiling model (from [Kommajosyula, 2020]).

is the one of Cole [Cole, 1960], which is improved in the work of Basu [Basu et al., 2005] and in the MIT Boiling model. For the departure diameter D_{d}, typical correlations are the ones of [Tolubinsky and Kostanchuk, 1970] and [Kocamustafaogullari and Ishii, 1995]. The MITB improves on these correlations by using adimensional numbers that take into account flow parameters such as the liquid pressure, which is missing in the previous correlations. Two of the experimental correlations used in the MITB are presented in the next section.

5.2.2 Experimental correlations

Bubble departure diameter

In the MITB, a departure diameter correlation was formulated to include the effects of pressure, superheat, subcooling, and liquid velocity. This correlation is given by the following equation:

$$D_{\mathrm{d}} = 18.9 \times 10^{-6} \times \left(\frac{\Delta\rho}{\rho_g}\right)^{0.27} \times Ja_{\mathrm{sup}}^{0.75} \times (1 + Ja_{\mathrm{sub}})^{-0.3} \times u^{-0.26}, \tag{5.1}$$

where:

- $\Delta\rho = \rho_l - \rho_g$ is the difference of densities between the liquid and gas phase,
- $Ja_{\mathrm{sub}} = \frac{\rho_l c_p \Delta T_{\mathrm{sub}}}{\rho_g h_{fg}}$ is the subcooling Jacob number, with c_{pl} the specific heat of the liquid, h_{fg} the latent heat, $\Delta T_{\mathrm{sub}} = T_{\mathrm{sat}} - T_{\mathrm{bulk}}$ the subcooling,
- $Ja_{\mathrm{sup}} = \frac{\Delta T_{\mathrm{sup}}}{\Delta T_{\mathrm{sub}}} Ja_{\mathrm{sub}}$ is the superheat Jacob number, with $\Delta T_{\mathrm{sup}} = T_{\mathrm{w}} - T_{\mathrm{sat}}$ the wall superheat,
- u is the inlet liquid velocity.

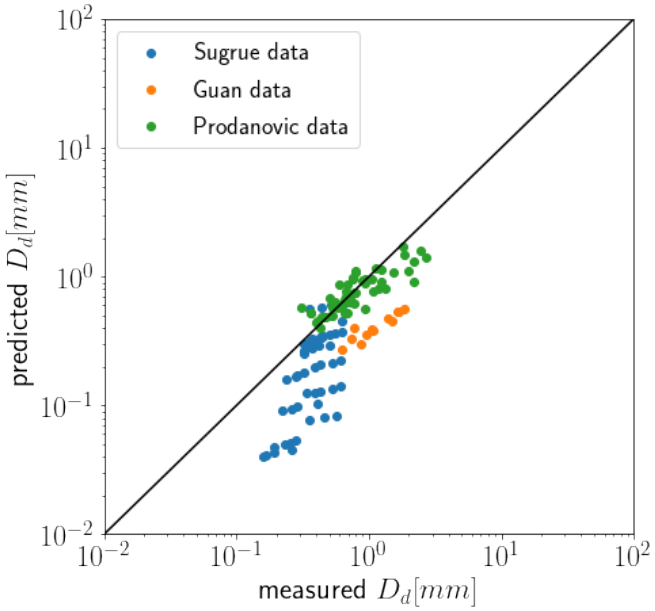


Figure 5.4: Predictions of the departure diameter correlation used in the MITB against experimental data.

The coefficients in the correlation were calculated by Kommajosyula using a least-squares fit done on diameter measurements obtained from 7 different experimental databases, the ones of [Prodanovic et al., 2002], [Situ et al., 2004], [Basu et al., 2005], [Sugrue et al., 2014], [Guan et al., 2016], [Yoo et al., 2016] and [Richenderfer et al., 2018]. Three of these databases are used for verification in this thesis: the ones of Sugrue, Guan, and Prodanovic. This data is used to evaluate the performance of the correlation used in the MITB.

On figure 5.4 are plotted the departure diameter predictions using correlation (5.1) against the experimental data. The correlation is mostly accurate on the Prodanovic data but distinctly underpredicts the Sugrue and Guan data. Improving the correlation with calibration with the inclusion of model error will give new predictions with credible intervals that will better represent the data.

Contact angle to compute the density of nucleation sites

A second experimental correlation used in the MITB is the Hibiki-Ishii correlation [Hibiki and Ishii, 2003], used to determine the density of active nucleation sites at the wall. Its expression, rather complex, involves the use of the static contact angle θ_c, which is an influential parameter. It is a property of the liquid and the type of wall surface, notably its composition, rugosity, and oxidization level. A typical value for water on a stainless steel surface is $\theta_c = 40°$. However, there is relatively high uncertainty on this value in practice because it is tedious to determine the surface parameters with precision. Besides, it is known that the contact angle depends on additional variables such as the pressure and the liquid temperature [Song and Fan, 2021], reaching variations of $\pm 30°$ in

a controlled environment while remaining in the hydrophilic range $[0°, 90°]$ for water and stainless steel. It is mentioned as a significant perspective in [Kommajosyula, 2020] to be able to make contact angle measurements in high pressure/temperature environments and determine their effect on boiling.

Thus, the contact angle is a prime candidate for calibration because it is influential on the wall heat flux predictions. Besides, when dealing with the Kennel experiments presented in the next section, it is difficult to estimate precisely simply from the experimental protocol described by the authors, as it was not directly measured. Additionally, we can explore if it depends strongly on flow conditions (ex: liquid temperature or pressure) since we work with multiple experiments.

5.2.3 Kennel experiment and model predictions

The experiment of Kennel [Kennel, 1949] is a study on the boiling of a subcooled fluid in various conditions of flow rate, pressure, and temperature. The investigations bear on nucleate heat transfer, pressure drops, as well as CHF transitions and heat transfer in film boiling. The test section is a Pyrex tube of length 0.457m and inner diameter 18.5mm, containing a central rod heater. The liquid flows in the tube and boils at the heater surface. The test section is shown in figure 5.5, and photographs of types of boiling obtained were shown in figure 5.2. The rod heater is a stainless steel tube of length 95.2mm which is heated using electrical power.

In this experiment, the wall heat flux is measured using the value of the instantaneous consumption of electrical power. The wall temperature is obtained with five thermocouples located inside the rod heater. The pressure drop is estimated from the manometer clasps visible in figure 5.5.

The measurement uncertainty is estimated by the authors as negligible on the wall heat flux data, as they eliminated heat and current losses (see p. 90 of [Kennel, 1949]). Temperature measurements on the wall are also subject to some error that is controlled by the presence of two additional thermometers, and the accuracy claimed by the authors is a margin of $\pm 0.5°F$. Due to other uncertainties, namely the temperature difference between the inside and the outside of the heating rod and the uncertainties on flow conditions that influence the theoretical saturation temperature, we consider in this work a margin of $\pm 2°F$ which corresponds to $\pm 1.11°C$, a typical uncertainty bound in boiling studies. This margin will be taken as a 3σ credible interval on the measured wall superheat value.

We extract the data from the manuscript of Kommajosyula, who used the experimental runs in the thesis of Kennel to assess the predictive capabilities of the MITB. In total, the model is calibrated using 13 runs, in the conditions presented in table 5.1. Remark that a smaller Pyrex tube was used in the last three runs.

The MITB performance was tested on this data as shown in figure 5.6. It can reproduce most of the experimental points within the experimental error bars, although exhibiting difficulties notably on cases 6, 14, 15, 18, 21 and 22. Thus the goal of calibration is to improve the regular predictions with error bars so the model itself can explain

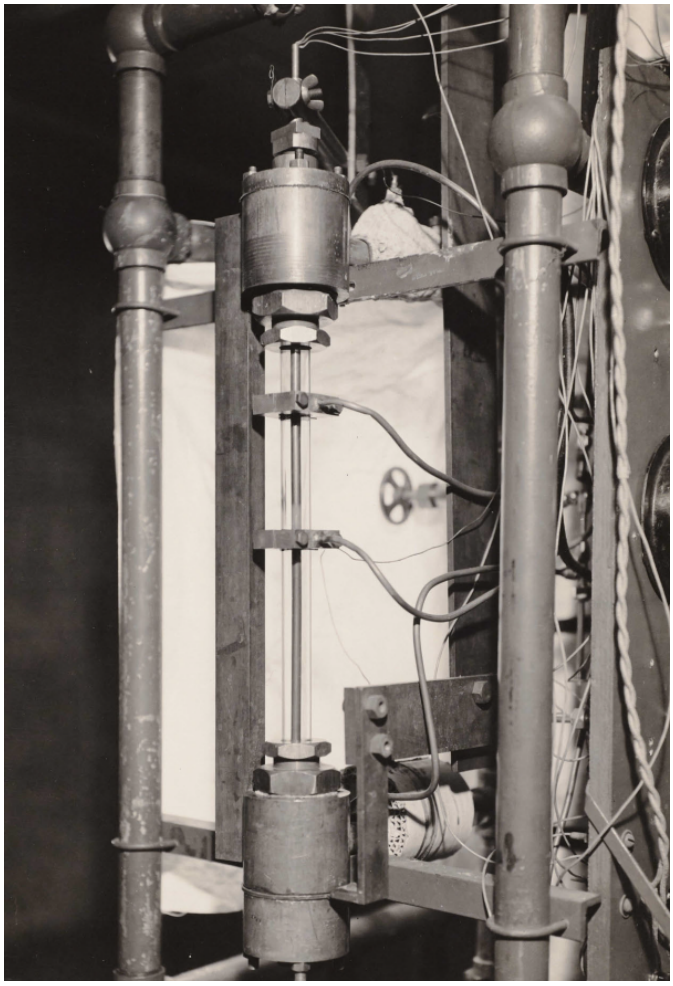


Figure 5.5: Setup of the Kennel experiment: a transparent Pyrex tube in which the liquid flows, with in the middle a rod heater that induces boiling. The rod heater is located between the two manometer clasps attached to the tube. At each extremity, the rod heater is soldered to copper elements (a tube above, a rod below), which passes the electrical current destined for heating.

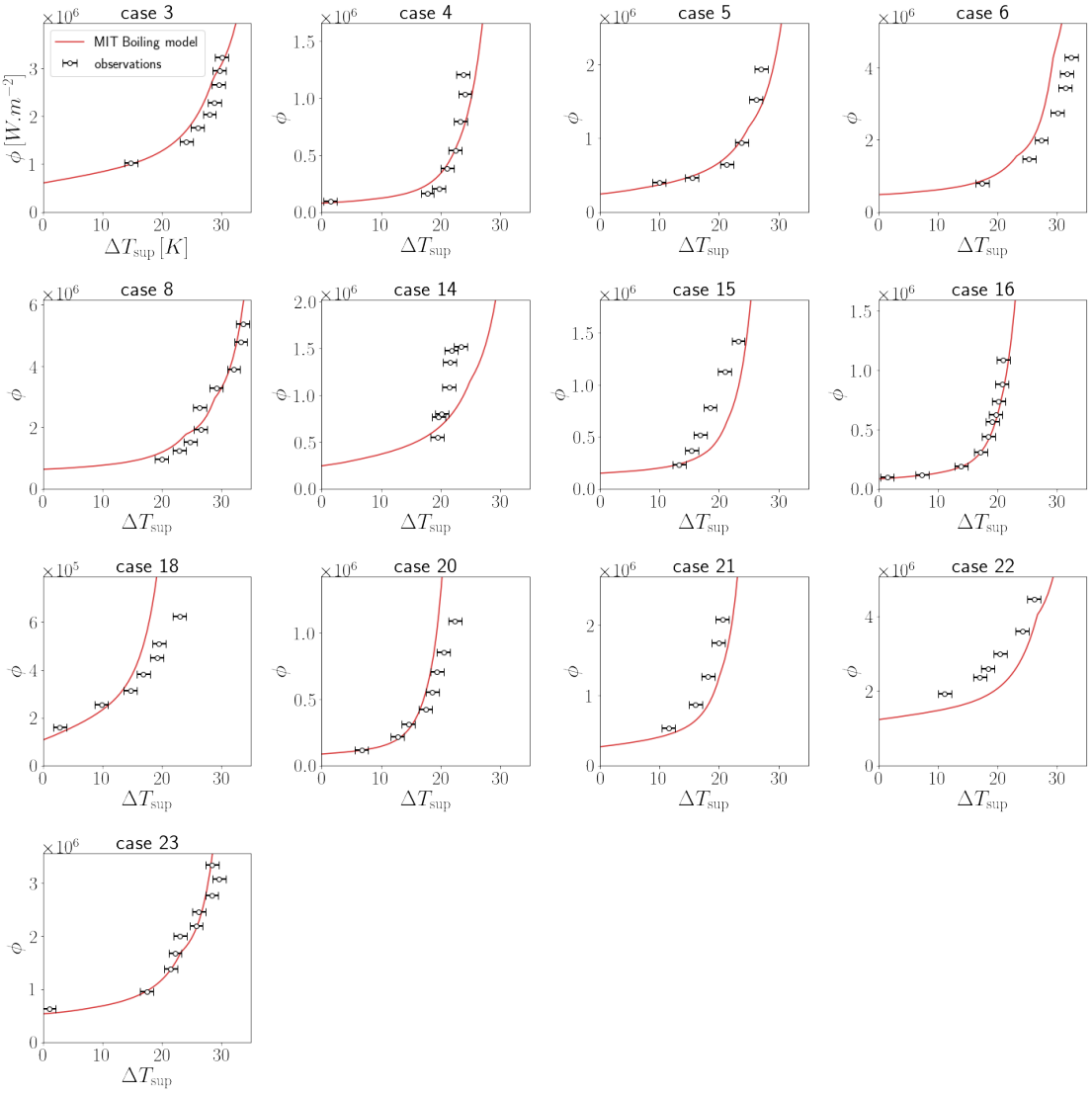


Figure 5.6: Predictions of the MITB on the Kennel data. Around the observations are shown the 95% confidence intervals due to the measurement error.

Case #	Pressure [bar]	Velocity u $[m.s^{-1}]$	Subcooling ΔT_{sub} [K]	Outer diameter [mm]	$\frac{\Delta\rho}{\rho_g}$	Ja_{sub}
3	2	3.66	27.8	19.56	853	45.4
4	2	0.3	27.8	19.56	853	45.4
5	2	1.22	27.8	19.56	853	45.4
6	4	1.22	55.5	19.56	446	48.9
8	4	1.22	83.3	19.56	454	74.2
14	2	1.22	27.8	19.56	853	45.3
15	4	0.3	55.5	19.56	446	48.9
16	4	0.3	27.8	19.56	437	24.2
18	4	1.22	11.1	19.56	430	9.6
20	6	0.3	27.8	19.56	294	16.8
21	6	1.22	27.8	19.56	294	16.8
22	4	3.35	55.5	10.67	446	48.9
23	4	1.22	55.5	10.67	446	48.9

Table 5.1: Operating conditions in the Kennel experiment, with corresponding values for the density ratio and subcooling Jakob number involved in the departure diameter correlation (5.1) (rightmost two columns). Ja_{sup} is a function of the wall superheat that varies, in each experiment, between 0K and 30K approximately.

more observations.

In each experimental run, two heat flux regimes can be distinguished: the first linear increase corresponds to the single-phase regime, and no boiling is observed. Then, as the wall temperature increases, the heat flux grows as a power curve: this corresponds to the establishment of the nucleate boiling regime. As mentioned before, in the experiments was also investigated the regime beyond nucleate boiling, past the CHF. These experimental points are not used in this study as the focus is on single-phase and nucleate boiling.

5.3 Calibration framework

5.3.1 Statistical assumptions and model error

In total, three parameters are chosen for calibration. The first parameter is the contact angle:

$$\theta_1 := \theta_{\text{c}}. \tag{5.2}$$

The other two are coefficients of the departure diameter correlation, so in the model, the diameter is computed with the following equation:

$$D_{\text{d}} = \theta_2 \times \left(\frac{\Delta\rho}{\rho_g}\right)^{0.27} \times Ja_{\text{sup}}^{\theta_3} \times (1 + Ja_{\text{sub}})^{-0.3} \times u^{-0.26}. \tag{5.3}$$

These two coefficients are chosen in the diameter correlation to avoid cofounding and low-information issues. In the Kennel experiments, only four different values of liquid

velocity are used, so it was considered insufficient to learn the exponent of u in the correlation. The same goes for the density ratio and the exponent of the $(1 + Ja_{\text{sub}})$ term. On the contrary, Ja_{sup} is directly proportional to the wall superheat, so it is expected to have a strong influence on the heat flux prediction.

The contact angle was chosen as 40° by Kommajosyula. We take a truncated normal prior distribution centered at this value, with a standard deviation 12° to allow reasonable variations. As the contact angle for hydrophilic surfaces should lie between 0° and 90°, we restrict the support of the prior between these values, which contain more than 99.7% of the prior mass. The other two parameters are also set a truncated normal prior distribution, with standard deviation at approximately one third of their nominal value, to stay close to the values estimated by Kommajosyula. Their support is bounded at low values by 0 since they are expected to be positive. For both parameters, the value 0 is further away than three standard deviations from the mean, so there is little probability mass lost in the truncation. The standard deviation is set so that the value "prior mean + 3 std" corresponds to twice the prior mean, which is coherent with the fact that the original correlation can underpredict the departure diameter by a factor 2 or 3 (see fig 5.4). Such informative priors can represent the seven sources of data used by Kommajosyula in building his correlation. All prior distributions are presented in table 5.2).

There are two approaches to parameter calibration in the case of multiple experiments:

- finding the best parameter values for each experiment, or
- finding the best global parameter value across all experiments.

Both these approaches will be explored in sections 5.4 and 5.5 respectively. Finding a global value has more appeal for confidence in using the calibrated model for future predictions. Finding best values individually is, in our opinion, more of a diagnosis for the presence of model error: if there is much variability between experiments, it is the structure of the model (or the choice of calibrated coefficients) that is at stake.

Accordingly, for the treatment of model error, there are two possible strategies:

- learn a unique model discrepancy term $z_{\boldsymbol{\theta}}$ across all experiments, or
- learn multiple model discrepancy terms $z_{\boldsymbol{\theta}}^{(i)}$, one for each experiment.

In this work, we explore only the second strategy, as the first would be inappropriate for two reasons: first, the heat flux values can be twice as much from one experiment to another, so we can expect the difference between observations and predictions to scale this way as well. Second, the onset of nucleate boiling (the point at which the power growth starts) varies across experiments, and if the model mispredicts it, the model discrepancies will not have the same shape across experiments. Consequently, there cannot be one single distribution of the model discrepancy term that could accommodate all experiments.

The $z_{\boldsymbol{\theta}}^{(i)}$ and their hyperparameters $\boldsymbol{\psi}^{(i)}$ are independent a priori, but in the case

where the value of $\boldsymbol{\theta}$ is shared across all experiments, they will not remain independent a posteriori. Each model discrepancy term is chosen with a zero mean and a Matern 5/2 kernel due to the expected regularity of the true process. Thus two hyperparameters are learned for each experiment, a standard deviation of the model error σ_i and a correlation length l_i. Their prior distributions are uniform with truncated support and bounds determined with respect to the heat flux's domain size and observed values (see table 5.2).

Variable	Prior	Support
θ_1	$\mathrm{p}(\theta_1) = \mathrm{N}_t(40, (12)^2)$	$]0, 90]$
θ_2	$\mathrm{p}(\theta_2) = \mathrm{N}_t(18.9 \times 10^{-6}, (5.7 \times 10^{-6})^2)$	$]0, +\infty[$
θ_3	$\mathrm{p}(\theta_3) = \mathrm{N}_t(0.75, (0.22)^2)$	$[0, +\infty[$
σ_i	$\mathrm{p}(\sigma_i) \propto 1$	$[10^4, 2{\times}10^6]$
l_i	$\mathrm{p}(l_i) \propto 1$	$[1, 25]$

Table 5.2: Prior distributions for model parameters and hyperparameters in the MITB calibration. The notation N_t corresponds to a normal distribution with truncated support.

5.3.2 Treatment of the measurement error

Let the MITB be represented with the function $f : T_\mathrm{w} \mapsto \phi$. As was mentioned in the description of the Kennel experiment, the experimental points were acquired by controlling the heat flux value with an electrical generator and making measurements of the wall temperature T_w. Thus, the experimental error lies in the temperature, which is an input of the model. Until now, we have only used techniques to include a measurement error on the output quantity in the calibration framework. Thus, additional treatment is required.

One possibility is to proceed with a standard calibration by using the inverse function of the MITB f^{-1}, which could be computed with root-finding numerical routine, which for a given heat flux would interrogate the MITB multiple times to find the corresponding superheat value. It exists because f is strictly monotonous and 1D. This fact would, however, multiply the computational cost of the method, so we adopt the Noisy Input Gaussian Processes (NIGPs) of McHutchon [Mchutchon and Rasmussen, 2011] which rely on a linearization of the target function (in the present case, the true process y) at the observed points, allowing to make a parallel between the input measurement variance and an output variance. We expand on this procedure below.

Suppose that in experiment $\#i$ we have $n^{(i)}$ measured temperatures and corresponding heat fluxes, noted $\{T_{\mathrm{w},j}^{(i)}, \phi_j^{(i)}\}_{j \leq n^{(i)}}$. Let the true wall temperatures, which would be measured if there was no measurement error, be noted $\{T_{\mathrm{w},j}^{(i),\mathrm{true}}\}_{j \leq n^{(i)}}$. The calibration equation reads, for $j \leq n^{(i)}$:

$$\phi_j^{(i)} = f(T_{\mathrm{w},j}^{(i),\mathrm{true}}, \boldsymbol{\theta}) + z_{\boldsymbol{\theta}}(T_{\mathrm{w},j}^{(i),\mathrm{true}}). \tag{5.4}$$

Suppose that the measurement error is gaussian homoscedastic, so that, for $j \leq n^{(i)}$, $T_{\mathrm{w},j}^{(i),\mathrm{true}} - T_{\mathrm{w},j}^{(i)} = \varepsilon_T \sim \mathrm{N}(0, \sigma_T^2)$. A first-order approximation of $f + z_{\boldsymbol{\theta}}$ gives:

$$(f + z_{\boldsymbol{\theta}})(T_{\mathrm{w},j}^{(i),\mathrm{true}}) \approx (f + z_{\boldsymbol{\theta}})(T_{\mathrm{w},j}^{(i)}) + \varepsilon_T * \frac{\partial (f + z_{\boldsymbol{\theta}})}{\partial T}(T_{\mathrm{w},j}^{(i)}). \tag{5.5}$$

Noting $d_j^{(i)} = \frac{\partial (f+z_{\boldsymbol{\theta}})}{\partial T}(T_{\mathrm{w},j}^{(i)})$, the calibration equation becomes:

$$\phi_j^{(i)} = f(T_{\mathrm{w},j}^{(i)}, \boldsymbol{\theta}) + z_{\boldsymbol{\theta}}(T_{\mathrm{w},j}^{(i)}) + \varepsilon_T * d_j^{(i)}. \tag{5.6}$$

In this new version of the equation, the measured points can be used to evaluate the MITB, with a heteroscedastic gaussian error on the prediction. Following the arguments of section 5.2.3, we consider a fixed measurement error with standard deviation $\sigma_x = 0.37°C$.

It remains to estimate the derivatives of $f + z_{\boldsymbol{\theta}}$ at the observed points. They are equal to the derivatives of the true process y, of which we have a good prior description, namely that it is strictly monotonous, regular, and fits the observations satisfactorily. The approach adopted here is to fit the observations with a parametric function and compute the derivatives once at the beginning of the calculation. Note that it would also be possible to work with the derivatives of the sequentially-fitted models $f + z_{\boldsymbol{\theta}}$, which would induce consequent supplementary computer cost, but could be an interesting alternative when little prior information about y is available. For more details about the online estimation of derivatives, see McHutchon [Mchutchon and Rasmussen, 2011].

In the present application, experimental conclusions have determined that the true process could be approximated with a linear function before boiling occurs and a power function. Another consideration is that the linear portion does not pass through the point $(0, 0)$, except when the liquid subcooling is $0K$, a case that is not present in the Kennel experiments. The derivatives of y are obtained by fitting through the observations the parametric function g given by:

$$g(x) = \begin{cases} c_1 + c_2 x & \text{if } x < c_3, \\ c_1 + c_2 x + (x - c_3)^{c_4} & \text{otherwise.} \end{cases} \tag{5.7}$$

Values of c_i are obtained with the least-squares fit, allowing to get the derivatives of g as estimates for the derivatives of y. This computation is done separately for each experimental case, once at the beginning of the code. A visual check is made to ensure that the parametric fit is appropriate, so the derivatives are correctly estimated, as seen in figure 5.7.

5.4 Calibration on individual experiments

In this section, the calibration is performed on each experiment independently. We compare the FMP calibration with the KOH calibration and the reference solution, which is the Bayes calibration. Due to the low number of unknowns, the results obtained with Bayes calibration are reliable and serve to compare the precision of FMP and KOH calibration.

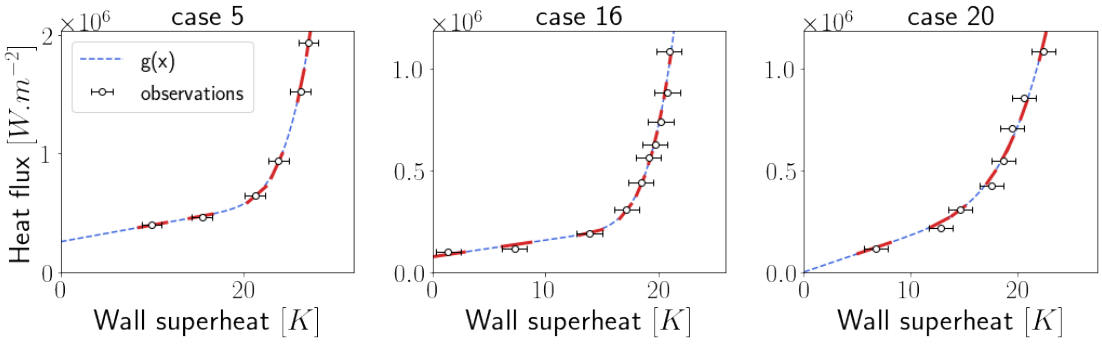


Figure 5.7: Least-squares fit of a parametric function g (blue, dashed) through the observations of each case to obtain the derivatives (red) at each observation point.

5.4.1 Likelihood function

Note $\mathbf{T}^{(i)} = (T^{(i)}_{\mathrm{w},1}, \dots, T^{(i)}_{\mathrm{w},n^{(i)}})^T$ and $\boldsymbol{\phi}^{(i)} = (\phi^{(i)}_1, \dots, \phi^{(i)}_{n^{(i)}})^T$ the experimental points of experiment $\#i$. Note also $\mathbf{f}^{(i)}_{\boldsymbol{\theta}} = (f(T^{(i)}_{\mathrm{w},1}, \boldsymbol{\theta}), \dots, f(T^{(i)}_{\mathrm{w},n^{(i)}}, \boldsymbol{\theta}))$ the predictions of the MITB, $\mathbf{D}^{(i)} = Diag(d^{(i)}_1, \dots, d^{(i)}_{n^{(i)}})$ the diagonal matrix of derivatives, and $\mathbf{C}_{\boldsymbol{\psi}^{(i)}}$ the covariance matrix of the model discrepancy.

The likelihood function for experiment $\#i$, noted $\mathrm{p}^{(i)}$, is given by:

$$\log \mathrm{p}^{(i)}(\boldsymbol{\phi}^{(i)}|\boldsymbol{\theta}, \boldsymbol{\psi}^{(i)}) = -\frac{n^{(i)}}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{C}_{\boldsymbol{\psi}^{(i)}} + \sigma_T^2\mathbf{D}^{(i)}) - \frac{1}{2}(\boldsymbol{\phi}^{(i)} - \mathbf{f}^{(i)}_{\boldsymbol{\theta}})^T(\mathbf{C}_{\boldsymbol{\psi}^{(i)}} + \sigma_T^2\mathbf{D}^{(i)})^{-1}(\boldsymbol{\phi}^{(i)} - \mathbf{f}^{(i)}_{\boldsymbol{\theta}}). \quad (5.8)$$

The prior distributions are the ones presented in section 5.3.1.

5.4.2 MCMC sampling of the posteriors

The posterior distributions are sampled using the M-H algorithm presented in section 4.1.2. For each experiment, a chain in dimension three is run for FMP and KOH, and dimension 5 for Bayes, so proper mixing can be achieved in all cases. Taking the example of experiment #3, the chains were 5×10^5 steps long for FMP and KOH, and 1.5×10^6 steps long for Bayes. The mixing qualities were evaluated with self-correlation diagnostics presented in section 4.1.4, obtaining mixing lengths of 18.5,15.4 and 80.5, respectively. The self-correlation plots are shown on figure 5.8.

For each chain, the parameter and hyperparameter posteriors are then estimated using a sample of size 1000, obtained by picking states at regular intervals from the list of visited states in the chain. The same sample is used to compute the posterior predictions and the associated credible intervals. The posterior summaries (mean and standard deviation) reported below are obtained using the entire visited states. Note that the Bayes calibration provides a sample of parameters and hyperparameters, whereas for KOH and FMP only a sample of parameters is obtained. For KOH, the hyperparameter

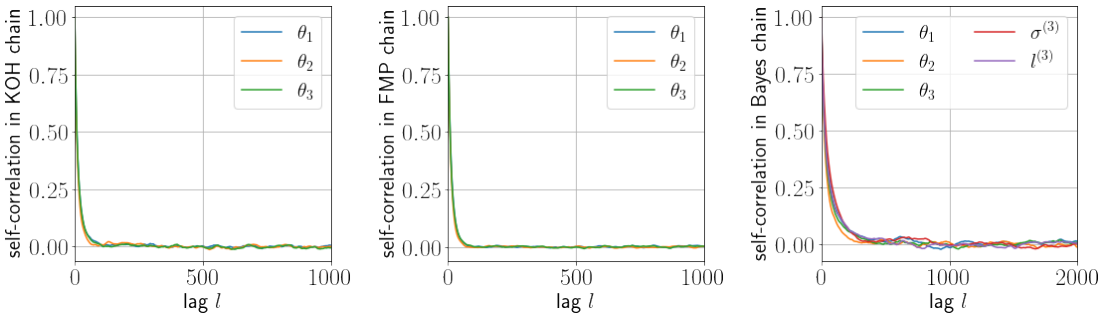


Figure 5.8: Self-correlation of the Markov chains for the KOH method (left), FMP (middle) and Bayes (right), for the individual calibration of experiment #3. Note that the x-axis on the Bayes plot does not have the same scale.

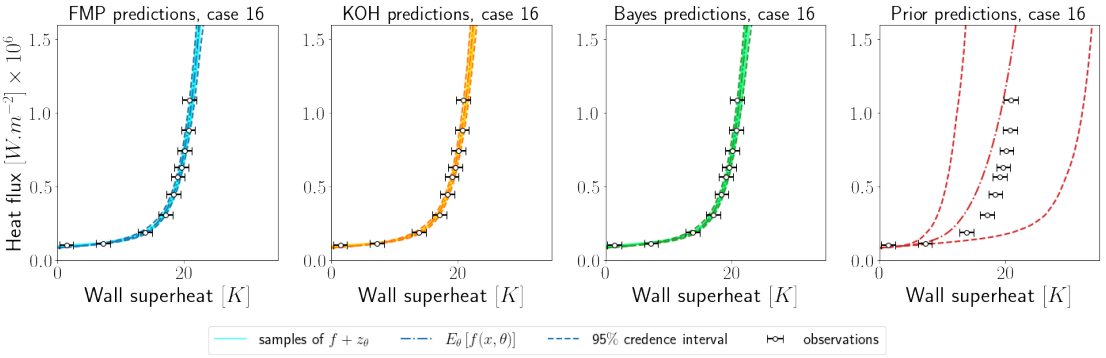
posterior is simply a Dirac distribution at their estimated value. In the FMP technique, a sample of hyperparameters is obtained by computing the optimal hyperparameters for each value in the parameter sample. This is coherent with the FMP approximation of the joint posterior.

The KOH estimation of hyperparameters is quick to obtain since it requires only a two-dimensional global optimisation per experiment, so conservatively, we set a maximum time of 20 seconds per optimisation. To accelerate the FMP method, we employ the optimal hyperparameter surrogates technique with an adaptive choice of training points presented in section 4.3. The surrogates are used as starting points for the FMP optimisations, which can then be quick local optimisations with a maximum time 10^{-4} seconds. For each experiment, the surrogates are built first using a space-filling nDoE of 100 points and are then enriched two times with 200 points obtained from successive MCMCs as described in the adaptive procedure. On each training point, a global FMP optimisation must be performed, and each is set to a maximum time of one second. Thus, the surrogate building phase requires 500 seconds of optimisation, and a cost in MCMCs which is small because the number of observations per experiment is low (see fig. 5.6, at most ten observations in experiments #8, #16 and #23).

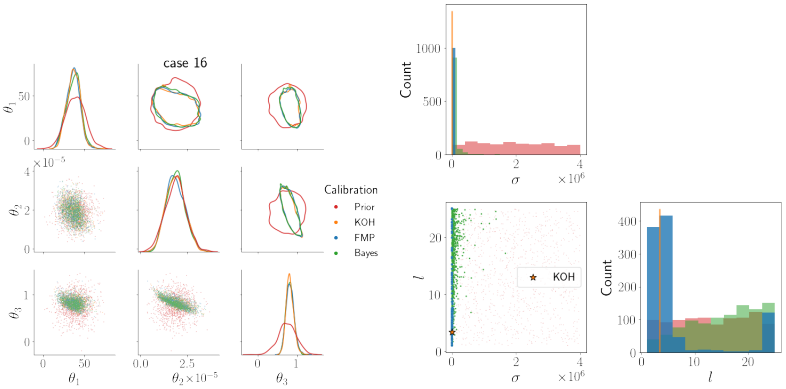
5.4.3 Posterior distributions

After calibration, the 13 experiments of the database can be classed into three groups for analysis:

- Group 1 (cases #6 And #16): The MITB can reproduce the observations with excellent precision for some parameter values.
- Group 2 (cases #3, #4, #5, #15, #18, #20, #21): some model predictions can explain the major part of the observations, but no particular parameter value stands out. There is a balance between model discrepancy and predictions.
- Group 3 (cases #8, #14, #22, #23): the MITB can not produce acceptable predictions. The model discrepancy is dominating.



(a) Posterior predictions



(b) Parameter posteriors (c) Hyperparameter posteriors

Figure 5.9: Calibration results on case #16 (group 1). The model discrepancy is noticeably low because the model predictions are accurate.
The color code, denoting the calibration technique used, is shared across the figures. The legend of (a) is shared with all its subplots, with unique colors for each calibration technique. The contour plots (upper diagonal on fig. (b)) represent 50% probability mass contours. The hyperparameter posteriors, represented with histograms, are not normalized for better visualisation.

We now proceed to the analysis of the results that concern each group.

The group 1 contains experiment #16 which is represented in figure 5.9. As could be seen already on the MITB predictions without calibration, the observations of case #16 were accurately reproduced by the model. Consequently, the posterior predictions are narrow in the calibration around the best model predictions. Regarding hyperparameters, the model error variance is narrow at its lowest value, and the correlation length value does not bear much influence: it is flat on the domain. Thus, the FMP and KOH approximations do not induce error on the parameters and predictions. Since the predictions before calibration are adequate for this case, the parameter posteriors are centered around the prior means, with a variance reduction for θ_1 and θ_3. In this group, all three calibration techniques globally agree.

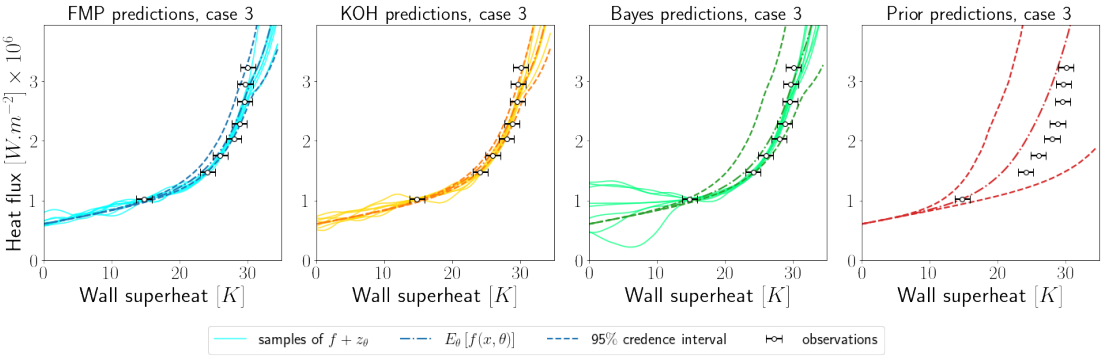
In group 2, we take the example of case #3, represented on figure 5.10. Here, some

model predictions are acceptable, but none is particularly excellent. Remark that an excellent model prediction would predict closely the experimental points: from a physicist's perspective, it would be acceptable if it lies in the experimental error bounds for each point. Still, the experimental error is gaussian in the chosen calibration framework so that a close fit would be heavily favored. Thus, σ is non-negligible, as seen on the corresponding Bayes posterior. The FMP posterior has a lower variance and correlation length as well. The parameter posteriors of KOH have slightly lighter tails than Bayes and FMP, which reflects on narrower credible intervals around the model predictions. Although the credible intervals of KOH can be considered as reasonable representations of the data, they do not reflect the whole range of predictions that Bayes accommodates, and FMP is closer in this regard. This behaviour is seen in almost all experiments of group 2.

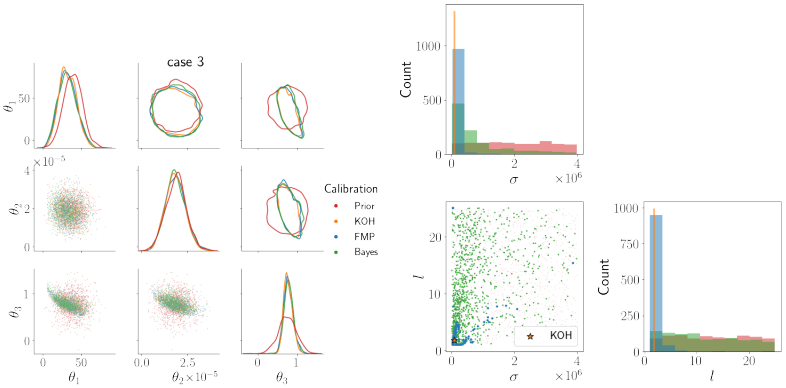
The group 3 includes experiments that are not well represented by the MITB, as case #23 shown on figure 5.10. Thus, the model discrepancy correction is strong, and most of the realizations of the true process (samples of $f + z_{\boldsymbol{\theta}}$) are not physically acceptable, as they show irregular oscillations or are non-monotonous. The prediction credible intervals provided by Bayes can include all observations within experimental error, and both FMP and KOH miss one observation. Here, the FMP and KOH approximations cause an underestimation of the posterior variance, showing the false confidence effect, which is stronger in the case of KOH.

From the global posterior summaries in table 5.3, a few global remarks about individual calibrations can be made:

- In all experiments, the parameter θ_3 is influent with a posterior distribution that differs noticeably from its prior. However, the posterior means can contradict one another: in case #5, its value is 0.884, and in case #18, the best value is 0.456, at a distance of more than three standard deviations. This point raises the question of whether a single value of θ_3 that accommodates all experimental cases can be found.
- Likewise, different optimal values of θ_1 and θ_2 are found, although these parameters are not influent in all cases. Besides, different optimal values remain close to one another, which is a good sign for calibration with all experiments.
- In the Bayes calibration, in most cases, the posterior distribution for l is primarily flat, the exception being for cases #8 and #23, which belong to group 3. On the contrary, the KOH estimation is generally at a low value of l, so the KOH trajectories can show spurious small oscillations that are an artifact of using a single value of l. These can also appear on the FMP trajectories or Bayes, but less frequently.
- The lower and upper bounds for the hyperparameters are sometimes met: either in the optimisations (KOH and FMP) or in the Bayes chain. They were chosen based on physical reasoning: too short of a correlation length would assimilate model error to measurement error, and one too big would accommodate model predictions far from the observations. Thus, the chosen bounds help keep the

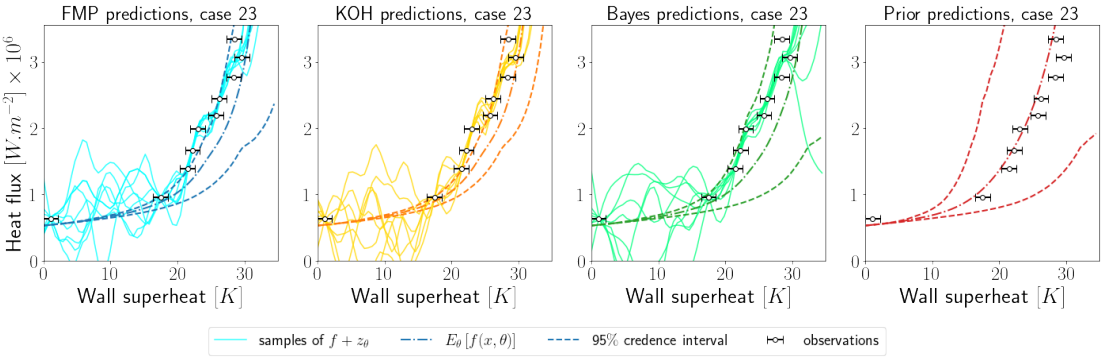


(a) Predictions #3

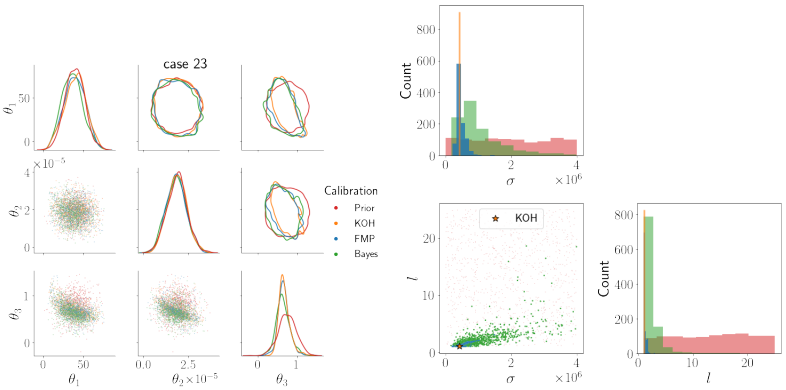


(b) Parameters #3

(c) Hyperparameters #3



(d) Predictions #23



(e) Parameters #23

(f) Hyperparameters #23

Figure 5.10: (a), (b), (c): calibration results on case #3 (group 2), where the model discrepancy and the model predictions are balanced. (d), (e), (e): calibration results on case #23 (group 3), where the model discrepancy dominates.

predictions in physically-acceptable ranges.

We also note that there are no obvious links between its posterior values and the four experimental conditions that vary in the Kennel cases in response to the previous inquiries about the contact angle. Thus, the individual calibration results do not reveal a potential new model for the contact angle.

5.5 Using multiple experiments

In this section, the calibration of model parameters is done using the data from a group of experiments. The central consideration in this section is that the dimension of the Bayes calibration with all experiments is 29, which causes mixing issues in the MCMC, so the method does not provide converged results. Thus, we first present calibration results using groups 1 and 2 of experiments, on which all methods are converged. Then, the FMP and KOH calibration results are compared when all experiments are used.

Another consideration is the definition of KOH hyperparameters: as maxima of the joint likelihood function, in theory, they should be computed using the data of all considered experiments and not determined on each experiment separately and used jointly. We refer to this method as the "KOH pooled" calibration and will compare its results to the KOH technique.

5.5.1 Group likelihood function and pooled KOH hyperparameters

Let $\mathbb{I}$ be a set of indexes of experimental cases in the Kennel database. The observations pertaining to this ensemble are noted $\boldsymbol{\phi}_{\mathbb{I}} = \{\boldsymbol{\phi}^{(i)}\}_{i\in\mathbb{I}}$, and the hyperparameters $\boldsymbol{\psi}_{\mathbb{I}} = \{\boldsymbol{\psi}^{(i)}\}_{i\in\mathbb{I}}$. The group likelihood function of the observations of the experiments that correspond to $\mathbb{I}$, noted $\mathrm{p}_{\mathbb{I}}$, is given by:

$$\log \mathrm{p}_{\mathbb{I}}(\boldsymbol{\phi}_{\mathbb{I}}|\boldsymbol{\theta},\boldsymbol{\psi}_{\mathbb{I}}) = \sum_{i\in\mathbb{I}} \log \mathrm{p}^{(i)}(\boldsymbol{\phi}^{(i)}|\boldsymbol{\theta},\boldsymbol{\psi}^{(i)}), \tag{5.9}$$

with $\mathrm{p}^{(i)}$ the likelihood for experiment $\#i$, given by equation (5.8).

Let $\boldsymbol{\psi}_{\mathrm{KOHp},\,\mathbb{I}}$ be the pooled KOH hyperparameters for set $\mathbb{I}$. Then, by definition,

$$\widehat{\boldsymbol{\psi}}_{\mathrm{KOHp},\,\mathbb{I}} = \arg\max_{\boldsymbol{\psi}_{\mathbb{I}}} \mathrm{p}(\boldsymbol{\psi}_{\mathbb{I}})\mathrm{p}(\boldsymbol{\phi}_{\mathbb{I}}|\boldsymbol{\psi}_{\mathbb{I}}) = \arg\max_{\boldsymbol{\psi}_{\mathbb{I}}} \mathrm{p}(\boldsymbol{\psi}_{\mathbb{I}}) \int_{\Theta} \mathrm{p}(\boldsymbol{\theta})\mathrm{p}_{\mathbb{I}}(\boldsymbol{\phi}_{\mathbb{I}}|\boldsymbol{\theta},\boldsymbol{\psi}_{\mathbb{I}})\,\mathrm{d}\boldsymbol{\theta}. \tag{5.10}$$

Two remarks can be made from this equation: first, their computation requires optimisation in the space of dimension $\dim(\boldsymbol{\psi}) \times \dim(\mathbb{I})$, which can cause a high numerical cost. Second, the pooled KOH hyperparameters have a priori nothing in common with KOH hyperparameters that would be determined on each experiment individually. This is an issue for practitioners, as acquiring new independent data would require the computing of new hyperparameters for the previous experiments.

Case (Group)	Method	θ_1	θ_2	θ_3	σ	l
#3 (2)	KOH	**33.0 ± 11.1**	1.8e-05 ± 5.3e-06	**0.769 ± 0.11**	**9.0e+04**	**1.8**
	FMP	**32.7 ± 11.6**	1.8e-05 ± 5.3e-06	**0.775 ± 0.12**	**9.9e+04 ± 1.7e+05**	**1.7 ± 1.7**
	Bayes	**32.3 ± 12.2**	1.9e-05 ± 5.4e-06	**0.795 ± 0.12**	**8.2e+05 ± 8.9e+05**	11.6 ± 6.9
#4 (2)	KOH	37.2 ± 11.2	1.9e-05 ± 5.2e-06	**0.828 ± 0.10**	**9.6e+04**	**1.5**
	FMP	37.4 ± 11.4	1.9e-05 ± 5.2e-06	**0.831 ± 0.11**	**1.2e+05 ± 8.3e+04**	**1.7 ± 0.8**
	Bayes	38.2 ± 11.3	1.9e-05 ± 5.3e-06	**0.851 ± 0.11**	**7.9e+05 ± 8.4e+05**	9.2 ± 6.6
#5 (2)	KOH	**30.4 ± 10.9**	1.9e-05 ± 5.2e-06	**0.870 ± 0.10**	**9.7e+04**	**2.5**
	FMP	**30.7 ± 12.5**	1.9e-05 ± 5.2e-06	**0.877 ± 0.11**	**1.3e+05 ± 1.2e+05**	**3.1 ± 1.9**
	Bayes	**32.0 ± 12.2**	1.9e-05 ± 5.2e-06	**0.884 ± 0.11**	**8.4e+05 ± 8.1e+05**	12.6 ± 6.4
#6 (1)	KOH	**36.9 ± 8.2**	1.9e-05 ± 5.0e-06	**0.645 ± 0.08**	**1.0e+04**	**25.0**
	FMP	**35.5 ± 10.7**	1.9e-05 ± 5.0e-06	**0.659 ± 0.10**	**5.6e+04 ± 1.3e+05**	**12.5 ± 10.7**
	Bayes	**32.2 ± 11.8**	1.9e-05 ± 5.1e-06	**0.684 ± 0.12**	**5.6e+05 ± 7.7e+05**	13.8 ± 6.7
#8 (3)	KOH	**39.4 ± 11.2**	1.9e-05 ± 5.4e-06	**0.758 ± 0.19**	**3.1e+06**	**1.0**
	FMP	**37.2 ± 11.1**	1.9e-05 ± 5.4e-06	**0.771 ± 0.18**	**2.0e+06 ± 1.4e+06**	**1.1 ± 0.9**
	Bayes	**34.3 ± 11.9**	1.9e-05 ± 5.4e-06	**0.794 ± 0.18**	**1.8e+06 ± 1.2e+06**	**3.8 ± 4.0**
#14 (3)	KOH	40.2 ± 11.4	1.9e-05 ± 5.4e-06	**0.781 ± 0.17**	**4.6e+05**	**1.5**
	FMP	40.4 ± 11.3	2.0e-05 ± 5.4e-06	**0.826 ± 0.19**	**4.6e+05 ± 2.3e+05**	**4.8 ± 7.9**
	Bayes	40.3 ± 10.7	2.0e-05 ± 5.4e-06	**0.892 ± 0.22**	**1.7e+06 ± 1.0e+06**	11.4 ± 7.2
#15 (2)	KOH	**36.6 ± 7.6**	**2.2e-05 ± 4.8e-06**	**0.903 ± 0.13**	**1.5e+05**	**2.4**
	FMP	**36.8 ± 7.0**	**2.2e-05 ± 5.1e-06**	**0.878 ± 0.15**	**1.5e+05 ± 1.9e+05**	**7.3 ± 9.4**
	Bayes	**37.3 ± 8.8**	**2.1e-05 ± 5.8e-06**	**0.818 ± 0.20**	**1.1e+06 ± 1.0e+06**	12.4 ± 6.5
#16 (1)	KOH	**36.3 ± 7.9**	1.9e-05 ± 5.0e-06	**0.821 ± 0.10**	**1.0e+04**	**3.4**
	FMP	**36.2 ± 8.2**	1.9e-05 ± 5.0e-06	**0.823 ± 0.10**	**1.1e+04 ± 3.9e+03**	**6.6 ± 7.0**
	Bayes	**37.1 ± 8.3**	1.9e-05 ± 4.8e-06	**0.816 ± 0.10**	**7.3e+04 ± 1.2e+05**	15.0 ± 6.3
#18 (2)	KOH	39.6 ± 11.2	**1.7e-05 ± 5.2e-06**	**0.484 ± 0.13**	**4.0e+04**	**2.0**
	FMP	38.3 ± 11.7	**1.6e-05 ± 5.3e-06**	**0.480 ± 0.14**	**5.2e+04 ± 2.2e+04**	**2.5 ± 1.1**
	Bayes	37.3 ± 12.1	**1.6e-05 ± 5.6e-06**	**0.456 ± 0.14**	**3.1e+05 ± 4.5e+05**	10.5 ± 6.7
#20 (2)	KOH	40.2 ± 11.5	1.8e-05 ± 5.1e-06	**0.571 ± 0.14**	**1.7e+05**	**1.8**
	FMP	40.6 ± 12.1	1.8e-05 ± 5.2e-06	**0.561 ± 0.15**	**1.9e+05 ± 9.5e+04**	**3.3 ± 2.2**
	Bayes	40.1 ± 12.5	1.7e-05 ± 5.3e-06	**0.539 ± 0.16**	**1.1e+06 ± 8.9e+05**	13.1 ± 6.3
#21 (2)	KOH	**47.8 ± 8.9**	**2.1e-05 ± 4.7e-06**	**0.787 ± 0.11**	**5.6e+04**	**25.0**
	FMP	**46.2 ± 9.9**	**2.1e-05 ± 4.8e-06**	**0.799 ± 0.12**	**9.0e+04 ± 8.0e+04**	**16.3 ± 10.0**
	Bayes	**44.4 ± 10.5**	**2.1e-05 ± 4.9e-06**	**0.809 ± 0.12**	**6.0e+05 ± 7.5e+05**	14.8 ± 6.4
#22 (3)	KOH	41.7 ± 11.0	1.9e-05 ± 5.1e-06	**0.754 ± 0.13**	**5.4e+05**	**5.4**
	FMP	41.2 ± 11.2	1.9e-05 ± 5.2e-06	**0.753 ± 0.15**	**7.1e+05 ± 3.2e+05**	**8.7 ± 4.4**
	Bayes	40.8 ± 11.3	1.9e-05 ± 5.2e-06	**0.719 ± 0.14**	**1.6e+06 ± 9.1e+05**	14.9 ± 6.0
#23 (3)	KOH	39.5 ± 12.9	1.9e-05 ± 5.2e-06	**0.679 ± 0.14**	**4.3e+05**	**1.1**
	FMP	38.9 ± 13.1	1.9e-05 ± 5.3e-06	**0.666 ± 0.16**	**4.6e+05 ± 1.4e+05**	**1.2 ± 0.3**
	Bayes	36.6 ± 12.6	1.8e-05 ± 5.4e-06	**0.650 ± 0.18**	**1.0e+06 ± 6.8e+05**	**2.3 ± 1.5**

Table 5.3: Posterior summaries for the calibration on each experiment independently. The statistics reported are the posterior mean with one standard deviation. In bold are the posteriors which noticeably differ from their prior distribution.

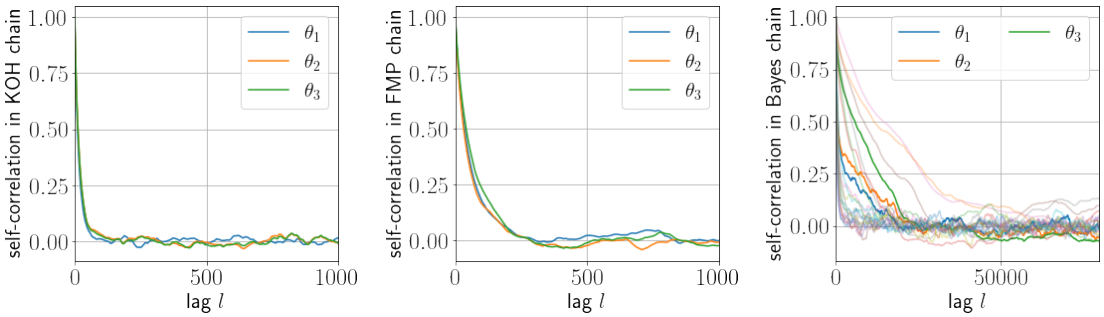


Figure 5.11: Self-correlation of the Markov chains for the KOH method (left), FMP (middle) and Bayes (right), for the calibration using experiments from group 1 and 2. The non-labelled curves in the right plot correspond to hyperparameters.

Thus, the original KOH definition leads to hyperparameters that would be valid on the entire group, despite causing difficulties in practice. None of these issues appears in the FMP calibration since optimal hyperparameters on each experiment remain optimal for the whole group due to the additive structure of the group likelihood function.

5.5.2 Using experiments from group 1 and 2

MCMC sampling

The sampling step for the KOH and FMP method is done the same way as the study on individual experiments, noting that the initial optimisation need not be rerun: the KOH hyperparameters and the FMP surrogates can be reused from the previous study due to the structure of the likelihood function. The sampling in the Bayes calibration requires more effort since, with the inclusion of 9 experiments, it is of dimension 21. Thus, mixing lengths of each method are 25, 94, and 23858, which can be interpreted as the maximal area under the curves in the self-correlation plots of figure 5.11. The convergence of the KOH pooled chain is similar to the one of the KOH chain.

The KOH and FMP chains are run with 5×10^5 steps to extract a sample of size 1000, so one every 500-th step, following their respective mixing lengths. To get the same quality of samples in the Bayes chain, we aim to extract one step every 50000-th step, which would require a total number of steps of 5×10^7. This is done by running for the Bayes calibration 10 chains, each of size 5×10^6, and extracting 100 points per chain. The starting points of each chain is drawn randomly in the space of unknowns, which helps diagnose convergence. Note that the mixing quality of the Bayes chains was obtained by taking as a starting proposal covariance matrix the sample covariance matrix of the FMP chain. This is expected to give good results since FMP is a correct approximation of Bayes, and without this step, the Bayes chain exhibits even poorer mixing. One of the ten chains is also started with the last state of the FMP chain.

The convergence of multiple chains can be assessed using the Gelman-Rubin factor of potential scale reduction [Gelman, 2014] that was presented in section 4.1.4. This factor compares in-between and within chains, variances to assess whether the chains should

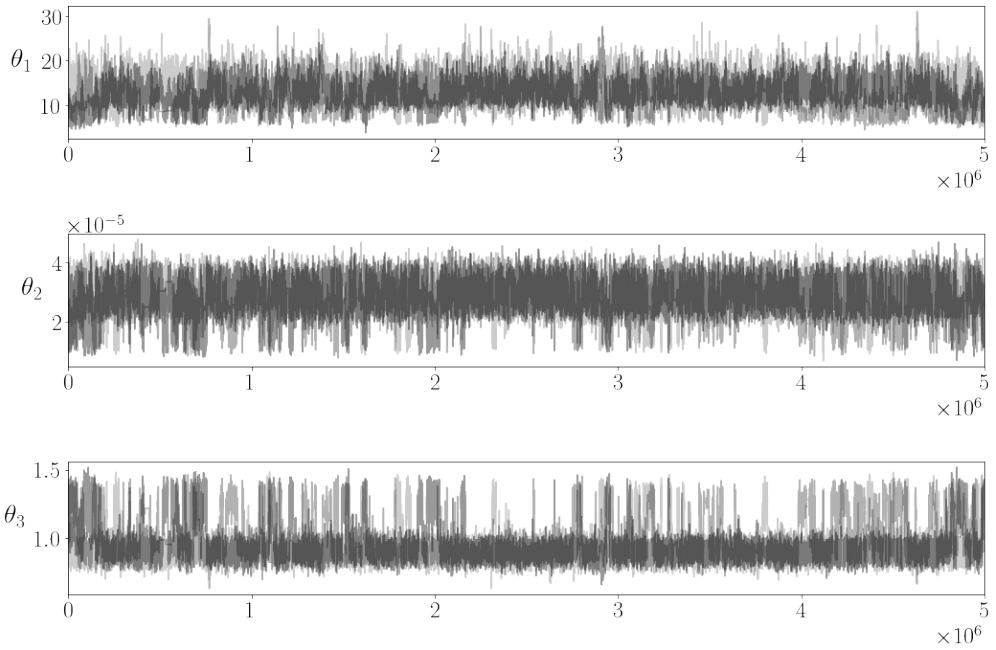


Figure 5.12: Trace plot for the three model parameters (up, middle and bottom) in the Bayes calibration of section 5.5.2. The three shades of gray correspond to the trajectories of three chains among the 10 that were run.

be simulated further. The authors suggest that obtaining a value $\hat{R} < 1.2$ suggests that convergence is attained. In our case, the factor is computed for each unknown and is always below 1.062, with the majority being below 1.01. A visual check can be made with the trace plots, shown in figure 5.12 for the parameters. It also helps confirm that the chains explore the same space regions despite starting from different points, which indicates the good convergence of the Bayes results.

The computation times for each method are presented in table 5.4. The initial optimisations of KOH are 20 seconds per experiment, and for the KOH pooled, it was conservatively set at 3600 seconds. The same goes for the FMP surrogate building, which required slightly more than 500 seconds for each experiment. The FMP surrogates can be reused with other experiments, such as in the next section, which is not the case with KOH hyperparameters. The Bayes chains were run longer than the others, but the computational time is relatively low because of all the out-of-bounds rejections of proposed steps. The FMP MCMC takes more time because each step involves a local optimisation. At first glance, the Bayes calibration seems to be more competitive in precision and computation time - however, it is only the case when good mixing is achieved. As will be seen in the next section, the Bayes calibration will not be accessible in higher dimensions.

Method	Steps			Total
	Initial optimisation	Surrogate building	MCMC	
KOH	180	0	15	195
KOH pooled	3600	0	20	3620
FMP	0	≈ 4600	1820	5420
Bayes	0	0	≈ 450	450

Table 5.4: Computation time (in seconds) for each calibration method with groups 1 and 2 of experiments.

Case	#3	#4	#5	#6	#15
σ	$1.9e+06$	$1.83e+05$	$1.8e+05$	$9.5e+05$	$3.7e+06$
l	15.0	12.1	13.6	5.2	20.1

Case	#16	#18	#20	#21
σ	$1.6e+06$	$4.8e+05$	$2.6e+06$	$1.6e+06$
l	23.5	1.8	9.3	8.2

Table 5.5: Point estimates of hyperparameters in the KOH pooled calibration with group 1 and 2 experiments.

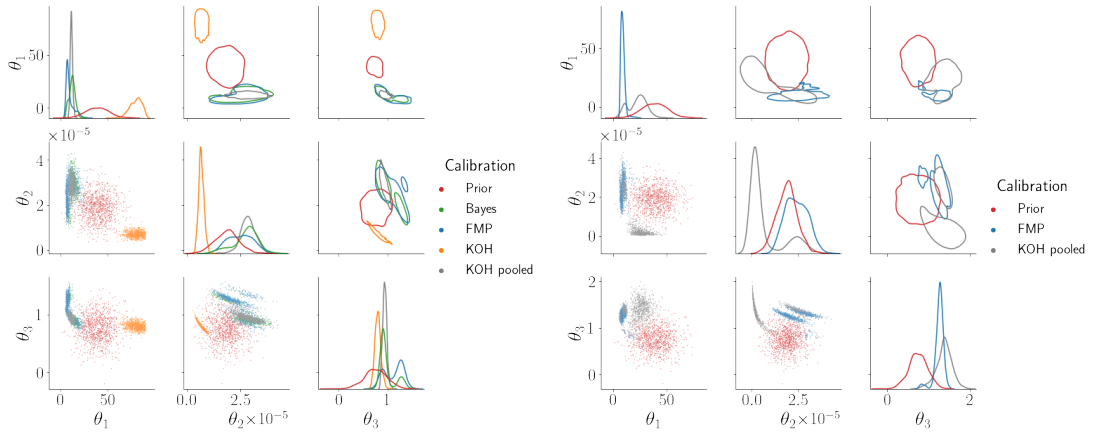
Posteriors and predictions

The parameter posteriors for all four calibration methods are presented in figure 5.13a. The true posterior distribution (Bayes) exhibits two modes. The true posterior plausible region is entirely missed by the original KOH calibration. The value of the contact angle predicted is at the upper bound (80°), whereas it should be at the lower bound ($\approx 10°$). The true marginal of θ_2 is also missed, but the prediction of θ_3 is acceptable since close to the true posterior mode. The KOH pooled method corrects the calibration, as it can capture the true posterior mode. It is, however, overconfident, as the variance is underpredicted in all dimensions, and the secondary mode in θ_3 is missed. The FMP method exhibits the best results of all three as its support covers exactly the posterior support. Both posterior modes are recovered: their relative importance is inverted, and the marginals are slightly distorted, but this bears no influence on the credible intervals of the posteriors and predictions.

The best parameter values are a low contact angle $\theta_1 \approx 10°$, a higher multiplicative coefficient for the departure diameter (θ_2), and also a higher exponent for the superheat Jakob number (θ_3). Note that the low value for the contact angle was not plausible in the individual calibrations, because the regularizing prior was used once per experiment. In contrast, in the group calibration the prior is used only once for all experiments.

The KOH pooled hyperparameters are presented in table 5.5. The value of σ is systematically greater than the ones determined on the individual calibrations, which shows that the pooled approach reduces the false confidence effect.

The posterior predictions obtained on four experiments are plotted in figure 5.14.



(a) With group 1 and group 2 experiments.

(b) With all experiments.

Figure 5.13: Parameter posteriors for the studies of section 5.5.2 (a) and section 5.5.3 (b).

Cases #6 and #16 that were extremely well-calibrated individually are well represented by the credible intervals on the model predictions in the Bayes and FMP methods, and the KOH pooled predictions are a bit off in case #6. The KOH predictions themselves are overconfident: case #6 seems to be the one that did determine for the KOH parameter posterior, as the observations are fitted almost perfectly, with credible intervals that are too restrictive. In comparison, in case #16, the predictions are now completely wrong as the optimal values for $\boldsymbol{\theta}$ do not match case #6. Besides, since a low model error was determined in the individual calibrations for this case, even the trajectories of $f + z_{\boldsymbol{\theta}}$ do not match the observations, which shows how mistaken the KOH calibration can be in this configuration.

The KOH pooled, FMP and Bayes methods provide more honest credible intervals that manage to capture the observations in most cases, but the KOH pooled calibration systematically underpredicts their size due to a more narrow parameter posterior. The FMP method exhibits robust, credible intervals that are a more accurate representation of uncertainty. However, none of the calibration methods achieves credible intervals around the model predictions that all capture the observations within the measurement uncertainty. For example, in case #3, they are captured by the Bayes and FMP method, but not all of them are captured by the KOH pooled method. In case #15, the three methods fail to capture the significant part of the observations. This trend shows that, when looking for parameter values that are optimal in some way over a group of experiments, we can expect some degradation of the individual behaviours.

5.5.3 Using all experiments

We now add to the pool of observations the experiments of group 3, representing four additional cases. As the original KOH calibration failed to produce accurate results in the previous section, it is excluded from the discussion in this section.

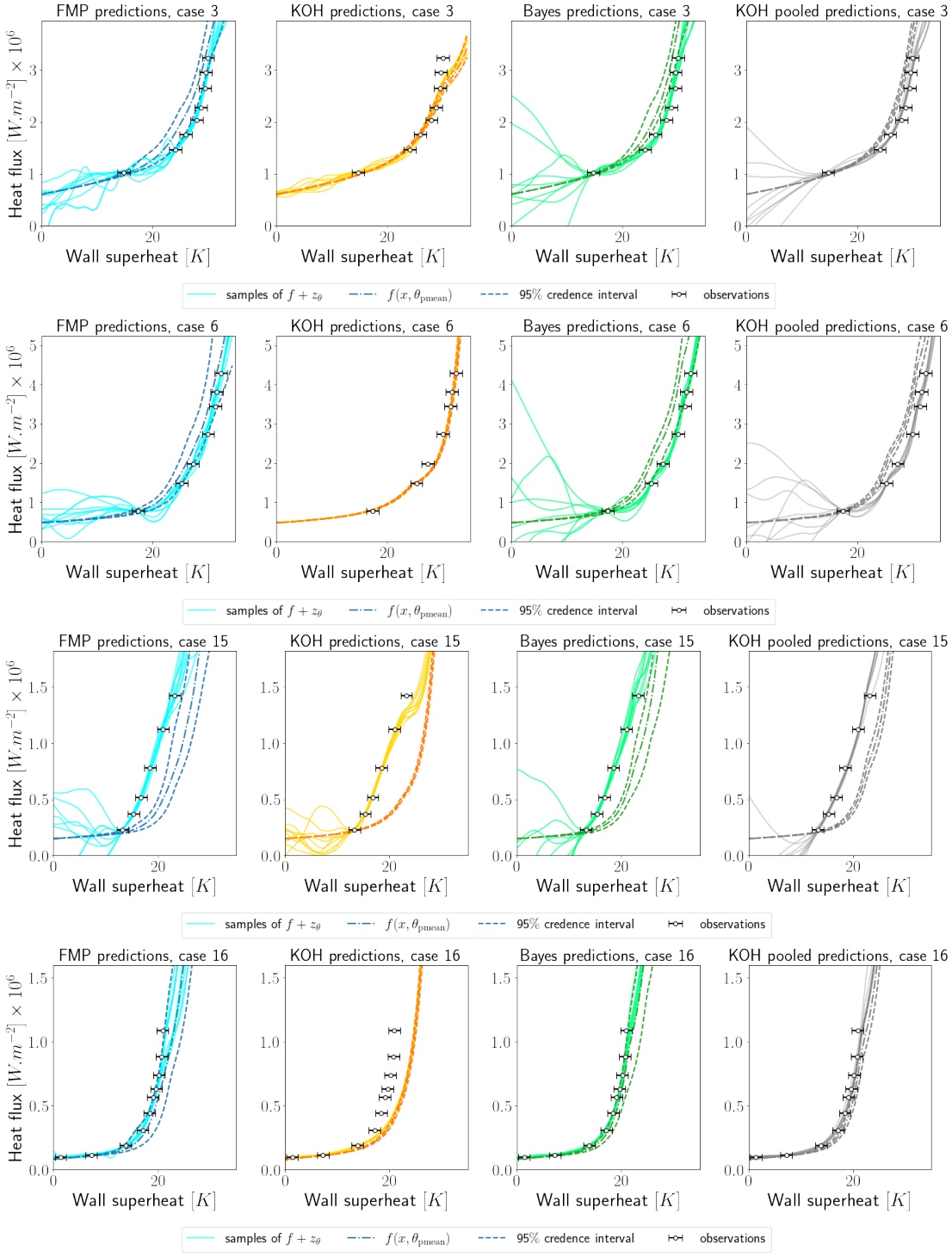


Figure 5.14: Posterior predictions on 4 experiments with each calibration technique when using the data of the experiments from groups 1 and 2.

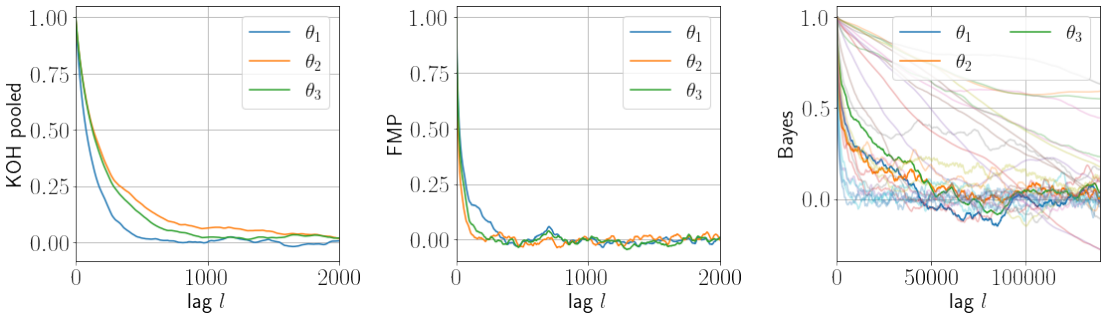


Figure 5.15: Self-correlation of the Markov chains for the KOH pooled method (left), FMP (middle) and Bayes (right), for the calibration using all experiments. The Bayes chain shows a lack of convergence.

MCMC sampling

The FMP chain is run in the same configuration as previously, with a mixing length of 49. The KOH pooled chain mixing was hampered by the exploration near the parameter space boundary, with a mixing length of 303. Consequently, the chain length was set to 1.5×10^6. The acceptance rate of both chains is around 5%. The Bayes chain, exhibits poor mixing with a mixing length that goes beyond 10^5. To run a converged Bayes chain using the techniques of the previous section would require considerably more computational power, so only the FMP and KOH pooled results are presented in this section. The self-correlation plots of the three methods are shown in figure 5.15. Note that the poor mixing in the Bayes chain mainly shows in the hyperparameters of experiments #4, #8, #20, #22, and #23, most of them which are in group 3.

Posteriors and predictions

The posterior distributions of parameters are shown in figure 5.13b. The added experiments have not influenced much the FMP posterior, which is similar to the one in the previous study, with less variance as more observations are used. On the contrary, the KOH pooled calibration explores a different region of the parameter space and provides posteriors that are significantly different from the previous study, as now low values of θ_2 are highly preferred, and the values of θ_1 and θ_2 are less certain. It is noteworthy that this time, the support of the FMP posterior is included in the KOH pooled posterior.

The hyperparameters estimated in the KOH pooled method are shown in table 5.6. Once again, the model error is estimated at a higher value for all experiments than in the previous section. Considering more experiments in the hyperparameter estimation phase tends to improve the conservativeness of this calibration technique to accommodate for more parameter values.

The predictions of the KOH pooled technique can now be considered as different interpretations of the data than the FMP calibration, as can be seen in the posterior predictions presented in figure 5.16. The credible intervals around the model predictions are generally more significant than the ones of FMP but do not help include additional

Case	#3	#4	#5	#6	#8	#14	#15
σ	$2.0e+06$	$2.3e+06$	$3.1e+06$	$1.3e+06$	$5.5e+05$	$4.2e+05$	$1.7e+06$
l	12.0	23.8	12.6	16.9	4.8	5.2	2.7

Case	#16	#18	#20	#21	#22	#23
σ	$3.1e+06$	$1.9e+06$	$1.6e+06$	$2.8e+06$	$2.4e+06$	$1.2e+06$
l	6.62	24.9	3.2	24.7	2.21	10.4

Table 5.6: Point estimates of hyperparameters in the KOH pooled calibration with all experiments.

observations, except in rare cases such as the first observations of case #3. In other cases such as #6 and #16, predictions are not improved either, and the added variance does not include the observations of case #15 that was poorly predicted previously. On the experiments of group 3 (right column of figure 5.16), the predictions of the model are closer to the observations in the FMP results. The predictions of the true process seems more physically acceptable for KOH in cases #23 and #8 (where the FMP interpretation is akin to observation error), but show overfitting in cases #14 and #22.

The observations of some cases in group 3, notably #8 and #23, exhibit variations that could mark a higher measurement error than what was reported, which could explain why the calibration is complex in these cases. Learning the measurement error in the framework could lead to a more accurate calibration. For cases #14 and #22, the model predictions are too different from the observations, so it is more likely that the model cannot reproduce these cases.

Overall, the addition of the experiments of group 3 did not affect much the results of the FMP calibration, but the KOH pooled results provided a different interpretation of the data, without the false confidence effect. As the FMP solution was close to the reference solution in the previous section, the FMP method exhibited its applicability and precision when dealing with a large space of unknowns in a practical case with multiple experiments.

5.6 Conclusion

This chapter presents the application of calibration with model error on an innovative heat flux partitioning model. This problem requires physical insights, notably for choosing prior assumptions, calibration framework, and treatment of measurement uncertainty. It is also numerically challenging due to the high number of unknowns required to include all experiments. In this application, the FMP approximation proved to be true to the reference solution, obtainable at low cost when the reference solution could not be computed, and robust to the inclusion of experiments that the MITB does not accurately represent.

The question is often asked whether it is better to find the best parameter values for each experiment or a single value representing the group as a whole. Our findings show that, by considering model error, it is possible to find global parameter values

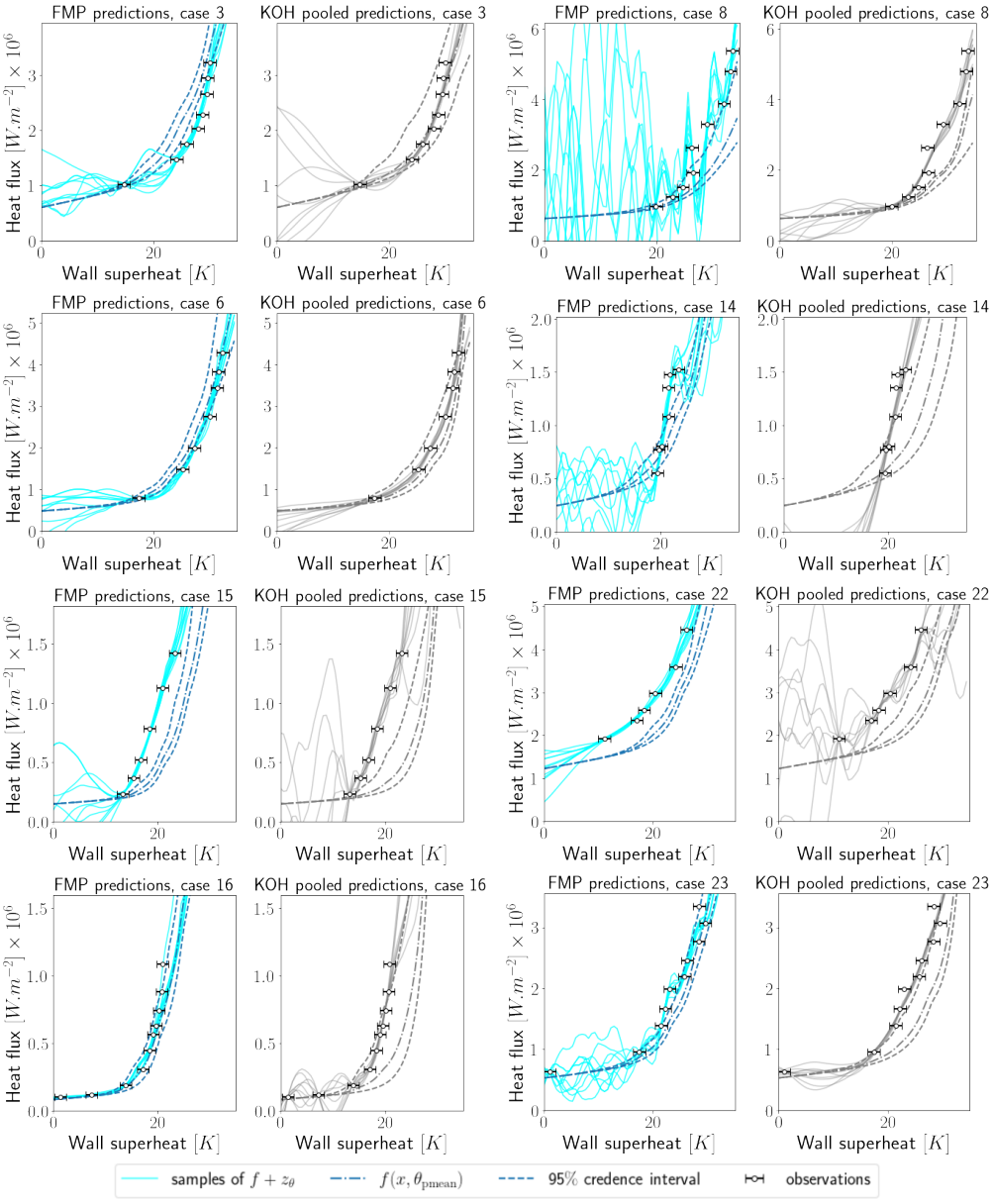


Figure 5.16: Posterior predictions with FMP and KOH pooled when using the data from all experiments. The right column shows the cases with predominance of model error.

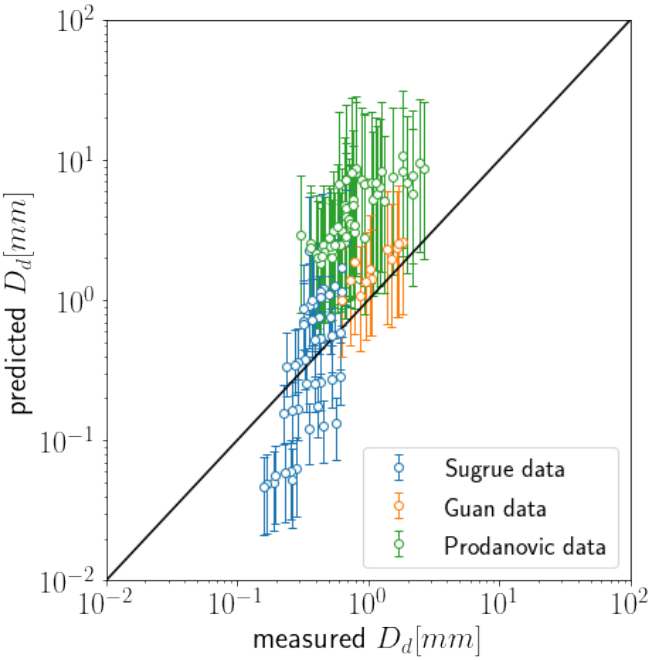


Figure 5.17: Predictions of the departure diameter correlation after calibration with the FMP method using experiments from groups 1 and 2.

associated with the uncertainty that provide correct predictions in most cases. These values correspond to a relatively low contact angle of around 10°, which could be an indicator that the surface of the heater used in the experiment was rougher than what was estimated by Kommajosyula. The departure diameter correlation with posterior values of coefficients can be plotted against the original data used for its determination, as is done in figure 5.17. The mean predictions are not as good a fit as the least-squares values (see fig. 5.4), but the attached 99% credible intervals give a more honest representation of the experimental points, especially on the Guan data. Probability-based predictions often offer the benefit of extending the validity range of models.

Chapter 6

Calibration of the NEPTUNE_CFD solver

Contents

This chapter consists of a complete calibration study of a model implemented in the solver NEPTUNE_CFD using data from the DEBORA experiments. The goal is to demonstrate that the techniques presented in the previous chapters can be used in a practical example to study the model error present in a model. A complete methodology is presented, from the choice of the relevant modifications of the models to the final calibration results. We study the influence of the choice to work with one or two quantities of interest and the choice of model discrepancy kernel (including a calibration without any model error inferred). On this application, we also compare the FMP, KOH and Bayes calibration results to prove the robust and accurate character of the FMP method.

6.1 Physical phenomenon: two-phase flow in a vertical pipe

In section 5.1 we have briefly described the mechanisms of heat transfer in boiling. The question at hand is to describe the type of flow that results from it, particularly in the case of water flowing upwards through a vertical pipe with heated walls, as represented in fig. 6.1. We work with an imposed liquid flow rate, so the regime is under forced convection. At the inlet of the pipe, the water is in a fully liquid state, and we suppose that its temperature is close enough to its saturation temperature that boiling will occur along the pipe, progressively replacing the liquid with vapor.

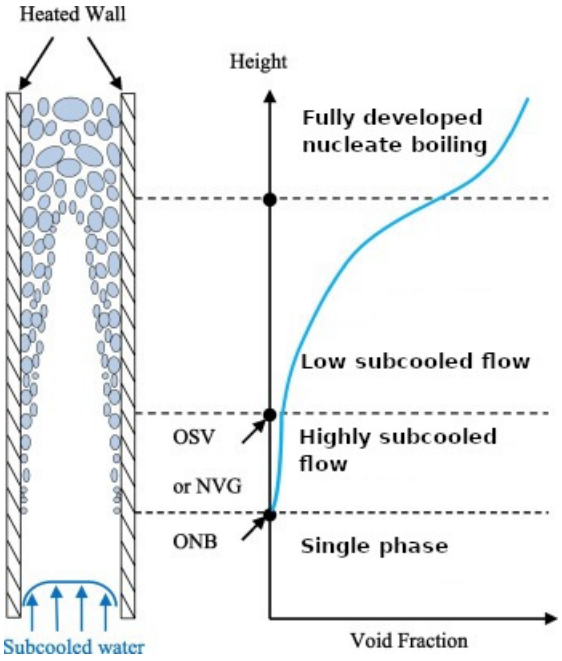


Figure 6.1: Boiling of water in a vertical pipe with forced convection, from [Ha et al., 2020].

Two additional states characterize the transition from single-phase flow to fully developed boiling. In all those states, we remain below the CHF, so still in the convective and nucleate boiling regimes of the heat transfer (according to fig. 5.1).

- In the single-phase flow, the water is subcooled (below its saturation temperature), and the liquid absorbs the heat transfer from the wall to raise its temperature along with the pipe height progressively.
- When the liquid temperature is high enough, still below its saturation temperature, the ONB occurs, and small bubbles form in the cavities of the wall. The bubbles size is limited due to condensation in the cold liquid. We refer to it as the Highly subcooled regime.
- As the void fraction and the liquid temperature increase, we get to the point of OSV or NVG, where the bubbles detach from the wall and migrate towards the

center of the channel. The void fraction increases significantly faster as more heat transfer goes towards ebullition. It is the Low subcooled regime.

- Eventually, the bubble interactions become increasingly important as they move in the section, break into smaller bubbles (breakup), or merge into bigger ones (coalescence). This is the fully developed Nucleate boiling state.

In the present application, we stop at the Nucleate boiling. Beyond that point, the CHF is reached, and the flow turns to the Slug Flow and Churn Flow regimes, with the merging of big bubbles to form pockets of vapor, then the vapor occupies the center of the channel, and only a liquid film remains at the walls, referred to as the Annular Flow. Eventually, a Mist Flow occurs when small drops of liquid are carried in a dominant gas phase. A full classification of two-phase flows can be found, for example, in [Ishii and Hibiki, 2006].

Two key quantities are crucial to accurately compute the void fraction in the channel: the prediction of OSV and the heat partitioning. The model of [Saha and Zuber, 1974] is a widely-used model for the OSV prediction, and for a more recent approach see [Ha et al., 2020]. Heat partitioning consists of the determination, in the Nucleate boiling state, of parts of heat transfer that go towards the liquid heating, the evaporation of the liquid, or the vapor heating. Such models can be either based on experimental correlations ([Jens and Lottes, 1951; Thorn et al., 1965]) or mechanistic ([Kurul and Podowski, 1990]). In Chapter 5 we considered a heat partitioning model that belongs to the second generation of mechanistic models [Kommajosyula, 2020]. On a larger scale, these models can be implemented in CFD solvers, which aim to simulate the three-dimensional flow. To this end, we present in the next section the NEPTUNE_CFD solver.

The physical models for boiling need to be supplied with experimental data to build and enrich the model or test its predictive capabilities. Some experimental databases where boiling phenomena are observed are [Roy et al., 1997; Lee et al., 2002; Situ et al., 2004; Yun et al., 2010], with the common feature that they were run at low pressures ($p < 30$ bar). In high pressure conditions, the experiments of [Estrada-Perez et al., 2009; INOUE et al., 1995; Garnier et al., 2001] are relevant for this application. In this work, we use part of the data that was acquired in the DEBORA experimental facility [Garnier et al., 2001], a boiling experiment that will be presented in more detail in section 6.3.1.

6.2 The NEPTUNE_CFD solver

6.2.1 The multiphase approach

The NEPTUNE_CFD code is a numerical solver for thermal-hydraulics simulations involving two-phase flows. It is a part of the joint research and development NEPTUNE program, initiated by French nuclear actors to build efficient simulation tools for flows occurring in PWR conditions, in nominal or incidental conditions. The consortium involved in this project is composed of EDF, CEA, IRSN and FRAMATOME (previously AREVA NP). NEPTUNE_CFD covers the finest simulation scale, which is the CFD

scale, where the goal is to reproduce local quantities, in opposition to simulation tools dedicated to representing an entire component of the circuit, or the circuit in its entirety. Some example applications for NEPTUNE_CFD are the 2D or 3D simulations of air-water, water-vapor, or stratified flows, in pipes, channels, or cavities. On the thermic side, the code can simulate boiling phenomena, condensation, liquid heating or cooling. Accurate dynamics models are also proposed to simulate jets, wall interactions, bubble migration, and flashing phenomena.

The method used for the simulation of flows is the multiphase approach of EDF, based on separate averaged Eulerian balance equations for the mass, momentum and energy for each phase. These equations are obtained by writing local balance equations for each phase and then averaging of the phase indicator. Averaged equations are closed with models for mass, momentum, and energy transfers between each phase. The turbulence is also treated in the liquid phase using specific RANS models. The numerical discretization is done by finite volumes, which allows using unstructured meshes for the calculation. The resolution in time is done with a first-order semi-implicit scheme in pressure, and velocity [EDF, 2020]. We now present the averaged balance equations and some closure models considered for the calibration.

Balance equations for two phases

Let the subscript k denote a phase, generally $k = l$ for liquid or $k = g$ for gas (note that the multiphase approach is relevant for more than two phases). The local balance equations are written for phase k then averaged over the temporal indicator of its presence, which results in the three averaged equations. The averaged quantities that are numerically resolved are $\alpha_k \in [0, 1]$ the average presence of phase k, ρ_k the average density, $\mathbf{v}_k$ the average velocity, p_k the average pressure, $\mathbf{T}_k$ the Reynolds stress tensor, and H_k the enthalpy.

The averaged equations for mass, momentum and energy that are solved in NEPTUNE_CFD write [Guelfi et al., 2007]:

$$\frac{\partial \alpha_k \rho_k}{\partial t} + \nabla \cdot (\alpha_k \rho_k \mathbf{v}_k) = \Gamma_k, \tag{6.1}$$

$$\frac{\partial \alpha_k \rho_k \mathbf{v}_k}{\partial t} + \nabla \cdot (\alpha_k \rho_k \mathbf{v}_k \mathbf{v}_k) = -\nabla(\alpha_k p_k) + \nabla \cdot [\alpha_k \mathbf{T}_k] + \mathbf{f}_{k,i} + \Gamma_k \mathbf{v}_{k,i} + \mathbf{f}_k^{ext}, \tag{6.2}$$

$$\begin{aligned}\frac{\partial \rho_k \alpha_k h_k}{\partial t} + \nabla \cdot (\rho_k \alpha_k h_k \mathbf{v}_k) = \nabla \cdot (\alpha_k \mathbf{T}_k \cdot \mathbf{v}_k - \alpha_k p_k \mathbf{v}_k - \alpha_k \mathbf{q}_k) \\ + q_{k,i} + \Gamma_k h_{k,i} + \mathbf{f}_{k,i} \cdot \mathbf{v}_{k,i} + \mathbf{f}_k^{ext} \cdot \mathbf{v}_k + q_k^{ext}.\end{aligned} \tag{6.3}$$

In equation (6.1), Γ_k is the mass transfer to phase k. In (6.2), $\mathbf{f}_{k,i}$ is the interfacial force applied on phase k and $\mathbf{v}_{k,i}$ the interfacial velocity. $\mathbf{f}_k^{ext}$ is the force applied on phase k that comes from an external source. In equation (6.3), $\mathbf{q}_k$ represents the heat transfer

internal to phase k, $q_{k,i}$ the heat source from the other phase, q_k^{ext} the heat source from external origins, and $h_{k,i}$ the interfacial enthalpy. A crucial assumption made in the multiphase formulation is that the pressure is identical for all phases, i.e. $p_k = p$ for all k. Note that these equations involve terms of exchange between phases, or external sources that required additional modelling, or "closure laws".

6.2.2 Closure models

Wall heat flux

In the DEBORA experiment, as the temperature of the liquid increases due to the heating at the walls, eventually boiling occurs when the nucleation sites at the wall are activated (ONB). In the code, these conditions are verified using Hsu's criterion [Hsu, 1962], which marks the transition between a purely convective heat flux into multiple fluxes, as given by the Kurul-Podowski extended model [Kurul and Podowski, 1990; Mañes et al., 2014]:

$$\phi_{\text{wall}} = f_l(\phi_{\text{cl}} + \phi_{\text{e}} + \phi_{\text{q}}) + (1 - f_l)\phi_{\text{cv}}. \tag{6.4}$$

The factor f_l represents the fraction of energy that goes into the liquid part of the mixture. The heat flux that goes to the liquid is decomposed into three parts:

- ϕ_{cl} is the convective flux that contributes to the heating of the liquid,
- ϕ_{e} is the evaporation flux which turns liquid into vapor,
- ϕ_{q} is the quenching flux, which is an accelerated convective heat flux due to the periodic arrival of cold liquid at the wall.

The heat flux that goes to the vapor, ϕ_{cv}, is purely convective.

Note that the wall heat flux partitioning model of chapter 5 is not implemented in NEPTUNE_CFD but could serve as an alternative to the Kurul-Podowski model.

Interfacial area transport equation

In bubbly regimes, as in the DEBORA experiment, it is crucial to accurately represent the average bubble size at each point of the flow. In regions with a larger quantity of interfaces, the heat and mass transfers are more important. Besides, different laws govern the dynamics of small and large bubbles. The bubble diameter can be calculated using the scalar field of interfacial area a_i, with units m^{-1}, which can be understood as the total area of the gas-liquid interface in a unit volume. The mean Sauter bubble diameter is defined as $D_m = 6\alpha/a_i$.

The evolution of interfacial area in the flow is represented with a transport equation, where the source terms are derived from models of bubble interactions. Concurrently, these interactions provide a second transport equation for the void fraction. These equations write:

$$\frac{\partial a_i}{\partial t} + \nabla \cdot (a_i v) = C_{\text{coal}} + C_{\text{bk}} + C_{\text{nucl},a_i} + C_{\text{mt},a_i} + C_{\text{comp},a_i}, \tag{6.5}$$

and:

$$\frac{\partial \alpha}{\partial t} + \nabla \cdot (\alpha v) = C_{\text{comp},\alpha} + C_{\text{mt},\alpha} + C_{\text{nucl},\alpha}. \tag{6.6}$$

The source terms in these equations represent the compressibility of phases, mass transfer between phases, nucleation (formation of bubbles at the wall by boiling), coalescence, and the breakup of bubbles.

These equations are solved at each time step in NEPTUNE_CFD, with a special numerical implementation: it is noted in [Ruyer and Seiler, 2009] that the transport equation that is numerically resolved is not directly equation (6.5), but rather the one that bears on $Xai = a_i/\alpha\rho_g$, thus the source terms in both equations (6.5) and (6.6) need to be computed. In the initial state, we start with a field α that was obtained from the alpha-pressure equilibrium (see [Guelfi et al., 2007]) and the field Xai from the previous time step. Then, the source terms are computed, and Xai is transported according to its equation. Finally, a_i is obtained by multiplication of these two fields.

From this procedure, we understand that a modification of the source terms results in a first-order effect on a_i but only a second-order effect on α. Thus, it will serve as a guide to propose in section 6.4.1 directions in which possible improvements of the models can be expected. In the remaining paragraphs of this section, we describe the multiple source terms that are obtained from the Ruyer-Seiler model [Ruyer and Seiler, 2009].

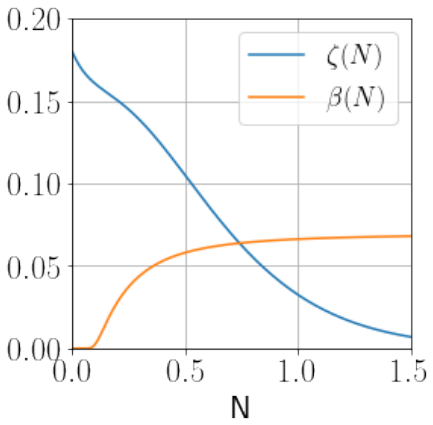


Figure 6.2: Experimental correlations for breakup and coalescence.

Coalescence and breakup

The coalescence of two bubbles is the merging of two bubbles into one with a bigger diameter, which reduces the number of interfaces while keeping the total gas volume constant. On the contrary, a bubble breakup into two or more bubbles will increase a_i with α constant. Following the Ruyer-Seiler model, the coalescence source term writes:

$$C_{\text{coal}} = -(\alpha\varepsilon)^{1/3} a_i^{3/5} \zeta(N), \tag{6.7}$$

with $N = D_{\mathrm{m}}^{5/6}\varepsilon^{1/3}\sqrt{\rho_l/\sigma}$ where ε is the turbulent dissipation rate, ρ_l the fluid density and σ the surface tension. The correlation, obtained from [Prince and Blanch, 1990], is given by $\zeta(N) = 0.18 * (1 + 2.75N + 12.9N^2) * \exp(-4.52N)$, as implemented in NEPTUNE_CFD 6.0.0 (see fig. 6.2).

The breakup source term is given by:

$$C_{\mathrm{bk}} = (\alpha\varepsilon)^{1/3} a_i^{3/5} \left(\frac{1-\alpha}{\alpha}\right) \beta(N), \tag{6.8}$$

where the correlation is derived from [Wu et al., 1998]: $\beta(N) = 0.06960.18 * (1 + 2.75N + 12.9N^2) * \exp(-4.52N)$ as implemented in NEPTUNE_CFD 6.0.0 (see fig. 6.2).

Nucleation

The wall boiling affects both void fraction (increase as more gas is generated) and interfacial area (increase as new interfaces form). The source terms are expressed as:

$$C_{\mathrm{nucl},\alpha} = \frac{\Gamma_{\mathrm{nuc}}}{\rho_v}, \qquad C_{\mathrm{nucl},a_i} = \frac{6\Gamma_{\mathrm{nuc}}}{D_d\rho_v}, \tag{6.9}$$

where $D_{\mathrm{d}} \neq D_{\mathrm{m}}$ is the diameter of the bubbles at the departure from the wall. It can be calculated following Ünal's model (see [Unal, 1976]).

Mass transfer and compressibility

The mass transfer source terms represent the effects of mass exchanges between the gas and liquid phases. On the void fraction, it amounts to:

$$C_{\mathrm{mt},\alpha} = \frac{\Gamma_p}{\rho_v}, \tag{6.10}$$

where Γ_p is the mass flux of vapor at the bubble interface, which can be obtained by using Ranz and Marshall modelling of the liquid heat flux at the interface (see [Ranz and Marshall, 1952] and equation 19 of [Ruyer and Seiler, 2009]). The complete expression of C_{mt,a_i} can be obtained with the same models (see [Ruyer and Seiler, 2009] eq. 20).

The compressibility terms are due to the variations of density of the gas phase in the flow. They are expressed as:

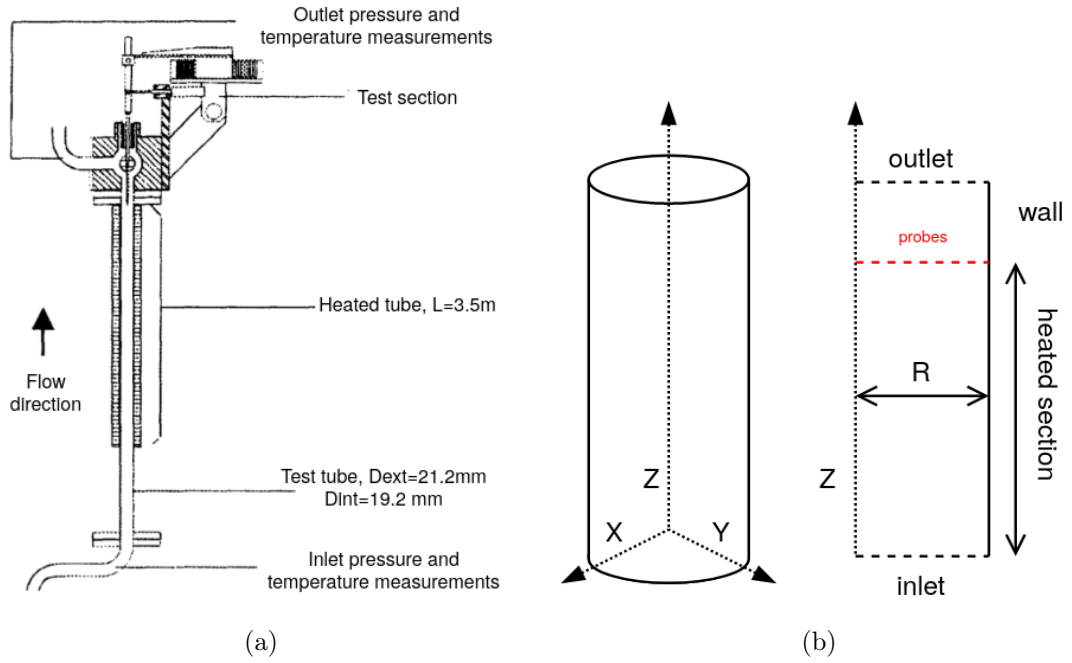
$$C_{\mathrm{comp},\alpha} = -\frac{\alpha}{\rho_v}\frac{\mathrm{d}\rho_v}{\mathrm{d}t}, \qquad C_{\mathrm{comp},a_i} = -\frac{2a_i}{3\rho_v}\frac{\mathrm{d}\rho_v}{\mathrm{d}t}. \tag{6.11}$$

6.3 Numerical simulation of the DEBORA experiments

6.3.1 Experimental setup

DEBORA is a thermal-hydraulics experimental facility that was exploited in the CEA center of Grenoble, France, for the study of boiling phenomena at a fine-scale [Garnier et al., 2001]. Following the experimental principle of scale similarities, the flow conditions

(pressure, flow rate, heating, temperature) are equivalent to the conditions encountered in PWRs, so that the boiling is representative of flows possibly occurring in these reactors. Because of the important vertical length of the tube in the installation, the transition from single-phase flow to bulk boiling can be reproduced in a single run.



(a) (b)

Figure 6.3: Schematics of the experiment (a) and internal geometry of the pipe (b).

The main body of the experiment is a vertical pipe of circular cross-section with internal diameter $D_{int} = 19.2mm$ and length $l = 3.485m$, with geometry represented in figure 6.3. The tube includes a heated portion over the major part of its length, which induces the boiling of the liquid. The working fluid, a coolant liquid (R12 or R134A), flows upwards through the channel. Four input parameters can be varied to produce different flow conditions: the working pressure, the inlet mass flow rate, the inlet liquid temperature, and the heating power conveyed to the fluid. Multiple experimental conditions have been explored through various campaigns cumulating over 400 experimental runs [Cubizolles, 1996; Manon, 2000; Garnier et al., 2001; Kledy, 2018], and they have been subsequently used for the validation of two-phase flows models [Ruyer and Seiler, 2009; Končar and Mavko, 2008; Krepper and Rzehak, 2011; Gueguen, 2013; Caner, 2020].

Measurements are obtained by optical probes located at the tube's cross-section at the end of the heated portion, at the outlet. Two types of probes were considered: single-tip probes allow for a measurement of the phase indicator, which gives a temporal signal from which it is possible to calculate void fraction, which is the proportion of time occupied by the liquid phase and the gas phase. We can access the average bubble diameter from this signal by assuming a theoretical velocity profile for the gas. The second type of optical probe is the dual-tip probe, which allows getting a measurement of the gas velocity to

get a more precise assessment of flow quantities such as the bubble distribution. Single-tip and dual-tip probes are mounted on moving support to acquire measurements in an entire radius of the cross-section; in particular, we can get observations at the wall or in the center of the channel. Additionally, the liquid temperature is measured along the heated portion with four thermocouple probes in some experimental runs. A detailed account of the measurement equipment, and the way the desired quantities are extracted from the phase indicator signal for both types of probes, can be found in [Garnier et al., 2001]. The observables that we will consider for study are:

- the void fraction α,
- the gas velocity V_g,
- the bubble diameter D equal to the mean Sauter diameter D_m,
- the liquid temperature at the outlet T,
- the interfacial area $a_i = 6\alpha/D$, which represents the total area of the gas-liquid interface in a unit volume of the mixture.

Note that the interfacial area, which directly relates to the void fraction and the bubble diameter, will be considered only for physical insight, but its measurements cannot be considered independent from the two observables in the calibration framework.

Experimental uncertainties of both inputs and observables are listed in table 6.1. The void fraction measurement is more precise than other observables because it is a first-order quantity on the phase indicator signal. Another point of interest is the 4% uncertainty on the heating power, taken from [Gueguen, 2013] and [Cubizolles, 1996]. In the original reference, it is noted that the heat flux is obtained by running an electrical current through the heating section and that the consumption of electrical power is measured with precision 1%. On the external side, the heating section is thermally insulated so that it is considered that the electrical power is converted into heating power for the liquid. We believe that uncertainty from [Gueguen, 2013] was obtained considering the variation of temperature of the wall in the heated section and potential heating losses, so we admit a 4% uncertainty on the heat flux transferred to the liquid in the remainder of this work.

6.3.2 Simulation of the A6 case

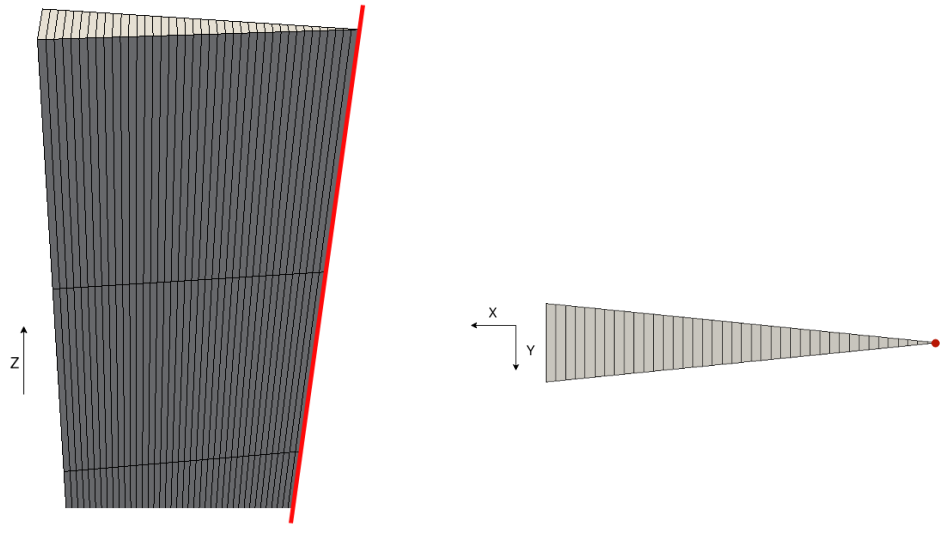
The configuration of the DEBORA experiment chosen for simulation is the A6 case, in which input conditions are summed up in table 6.2. This case was taken from the experimental campaign C2900 with the identification number 29G3P26W27Te64.8_1. In the C2900 campaign [Cubizolles, 1996] was used a single-tip probe, and no measurements of the liquid temperature were performed. This case was chosen because it belongs to the V&V database for NEPTUNE_CFD [EDF, 2019], and multiple boiling regimes are represented from the onset of nucleate boiling to $\alpha = 35\%$ at the end of the heating section. As observed in the simulation results, the void fraction in this calculation is slightly overestimated by NEPTUNE_CFD, which calls for calibration techniques to improve the model predictions.

Quantity	Unit	Uncertainty
Experimental inputs		
Outlet liquid pressure	bar	$\pm 50 \times 10^{-2}$
Inlet liquid temperature	$°C$	± 0.1
Mass flow rate	$kg.m^{-2}.s^{-1}$	$\pm 1\%$
Heating power	kW	$\pm 4\%$
Measured quantities		
Void fraction	-	$\pm 1\%$
Gas velocity	$m.s^{-1}$	$\pm 10\%$
Interfacial area	m^{-1}	$\pm 10\%$
Bubble diameter	m	$\pm 12\%$
Outlet liquid temperature	$°C$	± 0.2

Table 6.1: Experimental uncertainties on the DEBORA conditions on both experimental inputs and observables, from [Gueguen, 2013]. Uncertainties expressed in percentages are relative to the measured value.

Quantity	Unit	Value
Outlet liquid pressure	bar	26.17
Inlet liquid temperature	$°C$	65.58
Mass flow rate	$kg.m^{-2}.s^{-1}$	3062.2
Wall heat flux	$kW.m^{-2}$	12.88

Table 6.2: Experimental conditions for the DEBORA A6 case.



(a) (b)

Figure 6.4: 3D axisymmetric mesh for the simulation of the A6 case, viewed from the side (a) and the top (b). The axis of symmetry is highlighted in red.

A6 case is reproduced with a 3D-axisymmetric mesh representing the angular section of the pipe, extruded in its height, and the entire flow is obtained by symmetry around the principal axis of the tube. The refinement in the radial direction is 40 cells, and the height of the tube is divided into 400 cells, so the mesh comprises 16000 cells. The mesh is represented in figure 6.4. In a previous study of DEBORA simulations [Caner, 2020], a mesh refinement study was conducted, with the most refined mesh being composed of 64000 cells (refined twice in the axial and radial directions), and the simulation results for both meshes are similar. Thus, we consider that the current mesh is sufficiently refined for the calibration. Boundary conditions of inlet flow rate and temperature are imposed, and the pressure is prescribed at the outlet. The heat flux is set to match the experimental conditions at the walls. The models for dynamics (bubble forces and turbulence) used in the simulation are noted in table 6.3.

Quantity	**Model used**
Liquid turbulence	$R_{ij} - \varepsilon$ SSG
Vapor turbulence	none
Drag force	Ishii
Added mass	Zuber
Lift force	Tomiyama SMD
Turbulent dispersion	GTD
Wall force	Tomiyama

Table 6.3: Dynamics models used for the simulation of the DEBORA experiment.

The calculation is run with a fixed time step of $\mathrm{d}t = 1ms$, and according to the previous study of [Caner, 2020] the steady-state is assumed to be reached at $t_{max} = 23s$. The simulations are run in parallel over 20 cores on the COBALT cluster[1] of the French Centre de Calcul Recherche et Technologie (CCRT). In wall clock time, one simulation of the A6 case takes approximately 90 minutes. Numerical probes are located at the same height as the experimental probes so that after a simple interpolation, the simulation results can directly be compared to measurements.

The simulations are run for the A6 case using the prescribed models in NEPTUNE_CFD for DEBORA experiments. The results are compared to the measurements in figure 6.5, where r is the channel radius: $r = 0$m in the channel center and $r = 9.2 \times 10^{-3}$m at the wall. The numerical predictions are not satisfying with respect to the observations: the profile void fraction is generally overestimated (at the wall, around 40% instead of 30%), which gives an overestimation of the quantity of gas produced. The bubble diameter is correctly predicted at the wall but quickly decreases to a low value at the center, whereas in reality, the bubble size increases towards the center. The gas velocity profile is also overestimated in the middle of the channel.

[1] http://www-ccrt.cea.fr/fr/moyen_de_calcul/cobalt.htm

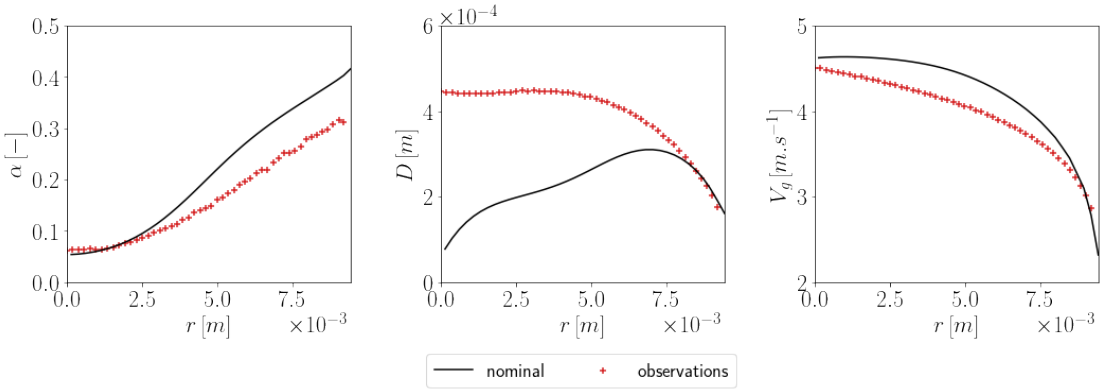


Figure 6.5: Predictions of void fraction, bubble diameter and gas velocity for the A6 case using the standard configuration of NEPTUNE_CFD.

6.4 Preliminary analysis of the simulations

6.4.1 Which models should be included in the calibration?

It is noted that Bayes' Theorem, and calibration techniques, can merely serve as a way of evaluating models and rating their performance in light of the experimental data. On the contrary, the proposal of new closure laws or modifying the existing ones bears on the physicist. In other words, the physicist asks the question, and Bayes' Theorem provides the answer. The interpretation of this answer can be made subjectively by the practitioner or following the more principled Decision Theory [Robert, 2007] that we do not develop here. Thus, the calibration framework must necessarily be grounded in physical principles, and it is the approach we follow here.

With this idea in mind, we formulate possible modifications of the simulation that could help reduce the mismatch between code predictions and observations in figure 6.5. We first take a look at leads to improve the void fraction predictions:

- setting a lower wall heat flux in the simulation would reduce the quantity of gas produced,
- setting a higher inlet pressure in the simulation would raise the saturation temperature of the liquid and lower the void fraction,
- setting a lower inlet liquid temperature would lower the void fraction,
- setting a higher inlet liquid flow rate would lower the void fraction,
- using a wall partitioning model which predicts a smaller proportion of evaporation heat flux would lower the void fraction.

Thus, a modification of the experimental inputs and the flux partitioning model should be explored as possible, leading to a reduction of the predicted void fraction. Note that the bubble diameter is correctly predicted where the void fraction is wrong (at the wall). Thus, as a first-order consideration, improving the bubble diameter will not significantly

affect the void fraction. The diameter influences the void fraction through intricate mechanisms such as the interfacial heat transfer and the lateral migration of bubbles that are too complex to be considered here.

For possible improvements of the bubble diameter predictions in the channel center, we formulate the following first-order modifications:

- setting a lower inlet liquid temperature would hinder the condensation of bubbles at the center and reduce the predicted diameter,
- diminishing the mass transfer due to condensation at the channel center would have the same effect,
- the phenomena involved in interfacial area transport (breakup and coalescence) could be inaccurately represented at the channel center. Specifically, the breakup would be overestimated, or the coalescence would be underestimated,
- the dynamics model for the radial migration of the bubbles could also play a role because if the big size bubbles would move faster to the channel center, it could raise the average diameter there.

For the gas velocity, the prediction error is relatively low (less than 10%) in the middle of the channel, and this quantity is not a significant output of the model compared to the void fraction. Consequently, we rather focus the sensibility study on the void fraction and the bubble diameter.

Another consideration that will drive our choice is the ease of implementing the proposed changes in the source code of NEPTUNE_CFD. In its quality of industrial code, numerical routines can be tricky to modify without introducing unwanted changes. Solving this issue would require a thorough practical understanding of the internal structure of the code, so in the thesis, we focused more on proposing a novel approach for sensitivity study based on simple considerations. As a consequence, we performed intrusive modifications with a straightforward implementation.

Note that a study on dynamics models for the lateral migration of bubbles has already been performed in [Caner, 2020], introducing a multiplicative coefficient in front of the lift force applied to bubbles, and the predictions were not significantly improved. Thus, we do not explore this assumption further in the present analysis. The other assumptions will be examined in the following, with a first OAT analysis of the three identified possible sources of error:

- the experimental uncertainties (section 6.4.2),
- the interfacial transport area model (section 6.4.3),
- the heat partitioning model (section 6.4.4)

6.4.2 OAT analysis of the experimental uncertainties

A OAT analysis consists in running a factorial nDoE (see section 2.3.1) by varying only one input while keeping all others at their nominal value. The value of the changed

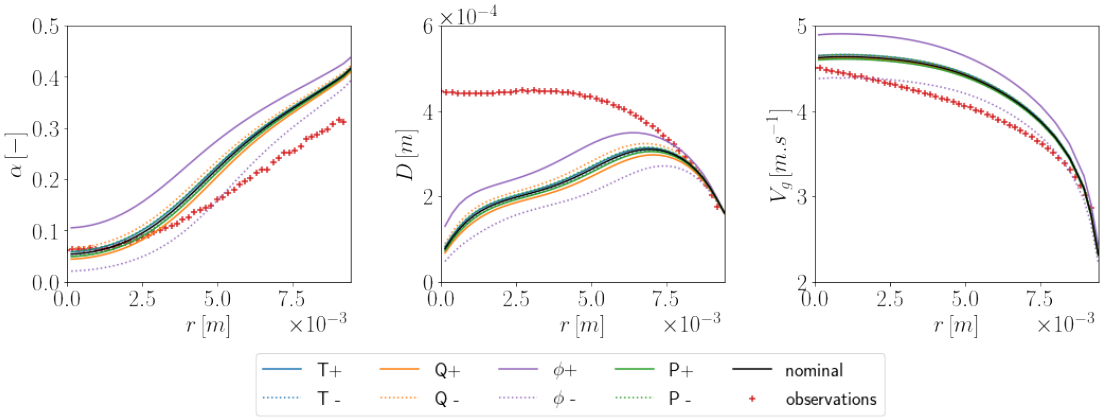


Figure 6.6: OAT analysis of the DEBORA simulations with varying experimental inputs: predictions of void fraction, bubble diameter and gas velocity.

input is generally taken at the boundaries of its domain so that minimal and maximal effects can be evaluated. Naturally, the second-order effect of varying two inputs at the same time cannot be estimated by this kind of design. Additionally, the intermediate variations between minimum and maximum are missed, and they might be important if the boundaries are too far away from the nominal value.

The nDoE is built on the four experimental inputs of the DEBORA experiments: the outlet liquid pressure, the inlet temperature, the mass flow rate and the wall heat flux. The maximal variation of each quantity is determined by the experimental uncertainty provided in the DEBORA studies, found in table 6.1. Both absolute and relative uncertainties can be treated with this method: the first by addition or subtraction to the nominal value, and the second by taking a percentage of the nominal value. In total, eight calculations are run.

The results of the simulations are plotted in figure 6.6. The effects of each input on the prediction of void fraction match the first-order effects that were explicit in the previous section. Predictions of bubble diameters are positively influenced by the wall heat flux, whereas the gas velocity is negatively affected by it, following the first-order effects. The predictions do not exhibit a strong sensitivity to the inlet temperature, flow rate and pressure due to their narrow range of variation. The sensitivity to the wall heat flux is, on the other hand, consequent, so that this variable will be considered in the calibration study.

6.4.3 OAT analysis of the interfacial area transport model

The interfacial area transport equation was presented in section 6.2.2. It is necessary to evaluate the level of detail of model modifications that we want to introduce to evaluate its sensitivity. Going from the highest level to the finest modifications, possible changes are:

1. At each time step, solve the regular transport equation for the interfacial area to

obtain the spatial field a_i, then correct it with a multiplicative coefficient, or by adding it to a spatial field that represents the error.

2. Modify the impact of the current models used to calculate the source terms in the transport equation (breakup, coalescence,...) by introducing multiplicative coefficients in front of them,

3. Modify the models directly for the source terms: introduce new correlation functions, allow a variation of some exponents, etc.

For the first type of modification, a global error coefficient independent of space and time would likely be not fine enough to represent the error on the interfacial area faithfully, so it would be more appropriate to consider a spatial and time-dependent field. This point poses the question of integrating this field to NEPTUNE_CFD, which would require in-depth source code modification. Since this approach is rather involved, we decided to try a more direct modification.

The third type of modification is the one that requires physical sense the most, as it concerns the direct expression of closure laws. We consider this kind of change to be on the frontier between proposing a new model and modifying the current one. Take the example of the expression for the coalescence coefficient given in eq. (6.7), which we recall here:

$$C_{coal} = -(\alpha\varepsilon)^{1/3} a_i^{3/5} \zeta(N).$$

The exponent of the turbulent dissipation rate ε is $1/3$ and was determined by sound physical considerations. For the sake of illustration, assume that changing its value to $1/2$ does create better predictions of the DEBORA experiments. Still, we would not advise changing it, except if it is derived from another physical consideration which is at least as sound as the original one. By doing so, we have effectively created another physical model that can be tested against experiments.

On the other hand, the expression of the correlation $\zeta(N)$ was assumed in its parametric form by the modeller and fitted against an experimental database. Keeping the same parametric form and performing calibration against our database would not inherently change the model used; instead, we would propose a new set of coefficients. We consider that changing the parametric form of the correlation would require a strong physical justification, as it would constitute, in fact, a new model.

We find appropriate the middle ground of introducing multiplicative coefficients in front of the source terms in the transport equation, for three reasons: first, for the breakup and coalescence, they play the same role as the constant coefficients in the expressions of the experimental correlations, so changing these values does not equate to creating a new model. Second, they can be directly interpreted as the measure of uncertainty of a given closure model. Third, their implementation in the source code is direct and thus less prone to errors.

Thus, we follow the approach of multiplicative coefficients in front of the source terms calculated by the closure models. A quick plot of the values of the terms in the standard

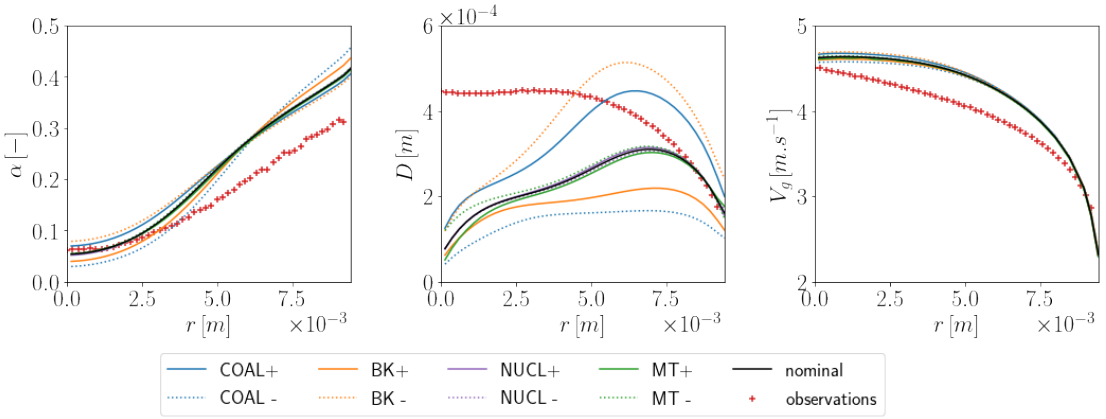


Figure 6.7: OAT analysis of the DEBORA simulations with multiplicative coefficients before the models for interfacial area transport: predictions of void fraction, bubble diameter and gas velocity.

calculation revealed that the compressibility term was negligible in front of the others, so it was not considered for modification. The multiplicative coefficients are noted $\{\theta_i\}_{i\leq 4}$, so the modified transport equations read:

$$\frac{\partial a_i}{\partial t} + \nabla \cdot (a_i v) = \theta_1 C_{\text{coal}} + \theta_2 C_{\text{bk}} + \theta_3 C_{\text{nucl},a_i} + \theta_4 C_{\text{mt},a_i} + C_{\text{comp},a_i}, \tag{6.12}$$

and:

$$\frac{\partial \alpha}{\partial t} + \nabla \cdot (\alpha v) = \theta_3 C_{\text{nucl},\alpha} + \theta_4 C_{\text{mt},\alpha} + C_{\text{comp},\alpha}. \tag{6.13}$$

To keep the modified values in a reasonable range of the values calculated in the Ruyer-Seiler model, we take the following domain of variation for the multiplicative coefficients: $\forall i,\ \theta_i \in [0.5, 2]$.

The coefficients are implemented in the source code of NEPTUNE_CFD, and the OAT analysis is performed with eight new simulations. The resulting predictions are plotted in figure 6.7, revealing that the proposed modifications significantly influence the predictions of bubble diameter. As expected, increasing the coalescence source term leads to a global increase of the bubble diameter, and the breakup term has the opposite effect. The reduction of the mass transfer term reduces the condensation effect in the center, which creates a favorable increase of the bubble diameter. The nucleation term does not bear much impact on the predictions.

These results indicate that the calibration of the closure models by the method of multiplicative coefficients will bring favorable results. Two things must be noted: first, they influence mainly the diameter predictions compared to the other two observables. The low impact on α can be related to the numerical implementation of the transport equation, as mentioned in section 6.2.2. Second, the bubble diameter remains underestimated at the channel center, and the observations do not lie in the range of the OAT. Consequently, no value of parameters contained in the variation range may be able to

get closer to the truth. The possibility still exists that, by varying multiple parameters simultaneously, the observations can be reached, but the only way to test this hypothesis is to run additional simulations.

6.4.4 Modification of the flux repartition model

To study the influence of the heat flux partitioning, we propose a modification of the extended Kurul-Podowski model. In a first step, this model is used to predict the partitioning of the heat flux to the liquid as $(\phi_{cl}, \phi_e, \phi_q)$ such that $\phi_l = \phi_{cl} + \phi_e + \phi_q$, at a given point on the wall. Then, this partition is changed into $\frac{\phi_l}{\phi_l'}(c * \phi_{cl}, \phi_e, \phi_q)$, with $\phi_l' = c * \phi_{cl}, \phi_e, \phi_q$. The effect is that the total heat flux to the liquid stays the same, but the relative importance of the convective heat flux is increased by a factor c.

Note that this modification of the partitioning model is limited to a simple readjustment step and does not hold much physical sense. The modified flux values are used directly in the models that require it (for example, the energy balance equations), but other values of the Kurul-Podowski model, such as the bubble departure diameter or the detachment frequency, are not impacted as they do not depend on the flux repartition. More complex modification strategies could have been explored, such as introducing new correlations for the active nucleation sites density or the quenching time, but let us remark that proposing such correlations would be akin to building a new physical model. The new repartition of fluxes is implemented in NEPTUNE_CFD, and three simulations

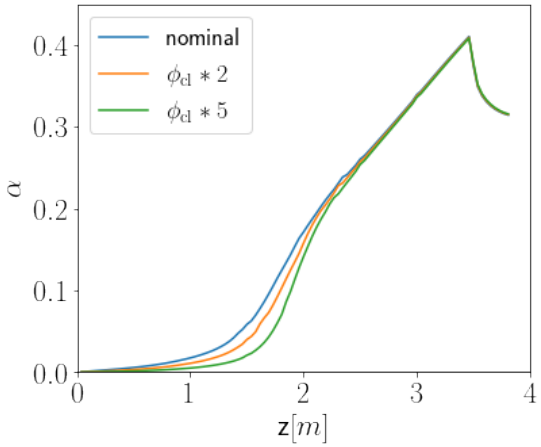


Figure 6.8: Void fraction at the wall as a function of the vertical abscissa in the channel for the standard calculation and the two calculations with the modified flux repartition.

of the A6 case are run: $c = 1$ (no modification of the model), $c = 2$ (slight increase of ϕ_{cl}) and $c = 5$ (strong increase of ϕ_{cl}). In figure 6.8 values of the void fraction are plotted, at the wall, as a function of the vertical abscissa z. At the entry of the heated section, no evaporation occurs, and the heat flux is only raising the liquid temperature. Along the heated section, boiling occurs, and the void fraction is created. In the computations where the evaporation flux is magnified, gas production occurs faster. As we progress in the heated section, all predictions converge to a single curve. The value $z = 3.485m$

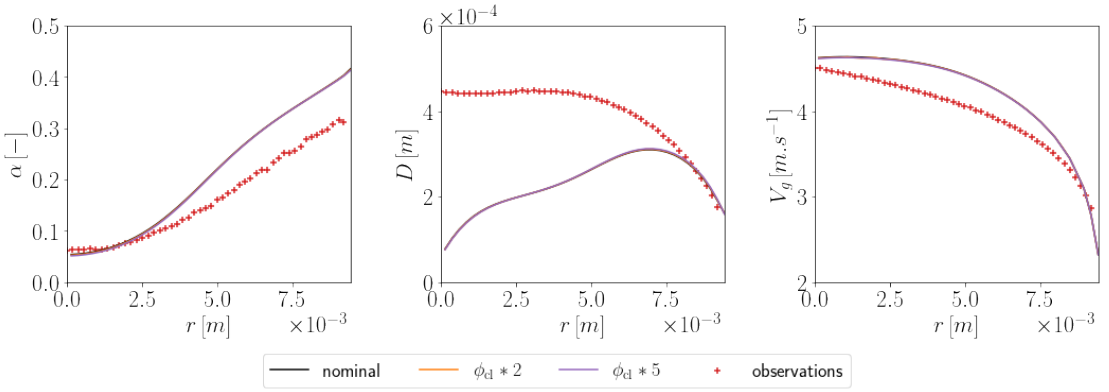


Figure 6.9: OAT analysis of the DEBORA simulations with modification of the flux repartition model: predictions of void fraction, bubble diameter and gas velocity.

is the end of the heated section, so the void fraction at the wall sharply decreases after. From this figure, we observe that the predictions at the probe locations are almost unchanged, as is plotted for verification on figure 6.9.

In conclusion, the modification of the flux partitioning model will not be included in the calibration framework since it does not influence the predictions of the desired quantities.

6.4.5 Partial conclusion

From the preliminary analysis, we retain the following insights:

- The void fraction predictions mainly depend on the experimental uncertainty related to the wall heat flux,
- The remaining experimental uncertainties are too low to bear an influence on predictions,
- The bubble diameter predictions mainly depend on the source terms from the interfacial area transport model,
- The various modifications proposed do not influence much the gas velocity predictions.

Thus, we have identified the variables that will be relevant to include in the calibration framework. The following section will detail this chosen framework and its application to the DEBORA experimental data.

6.5 Calibration study

6.5.1 Statistical assumptions and numerical considerations

In this chapter, five multiplicative coefficients are introduced in the source code:

- The four coefficients in the interfacial area transport equation presented in section 6.4.3, $\{\theta_i\}_{1\leq 4}$ with $\theta_i \in [0.5, 2]$,
- The coefficient θ_5 which is multiplied to the heat flux boundary condition imposed at the heated wall. We use $\theta_5 \in [0.9, 1.1]$.

Although the experimental uncertainty on the wall heat flux is claimed to be at most 4%, we allow a variation of 10% in our calculations. This is done so that the extremal predictions can include all the void fraction observations, giving a better chance at good predictions after calibration. A variation of 4% is still included in our range of variation so that if this value is deemed very plausible in light of the data, it will appear in the calibration results.

From the preliminary analysis, as experimental data we take observations of void fraction $\mathbf{y}_{\text{obs},1}$, and bubble diameter $\mathbf{y}_{\text{obs},2}$, that we consider independent. There are $n = 49$ scalar observation points $\mathbf{x}_{\text{obs}} = (x_{\text{obs},1}, \ldots, x_{\text{obs},n})^T$, which are the same for both quantities. The model predictions of each quantity are $f_1(x, \boldsymbol{\theta})$ and $f_2(x, \boldsymbol{\theta})$.

Two a priori independent model discrepancies $z_{2,\boldsymbol{\theta}}$ and $z_{2,\boldsymbol{\theta}}$ are chosen, with prior mean 0, and we note $c_{1,\boldsymbol{\psi}_1}$ and $c_{2,\boldsymbol{\psi}_2}$ their respective covariance functions indexed with their hyperparameters. The measurement errors for each variable, $\sigma_{1,\text{mes}}$ and $\sigma_{2,\text{mes}}$ are also independent a priori. We take a Matern 5/2 covariance function for each model discrepancy:

$$z_{1,\boldsymbol{\theta}} \sim GP(0, c_{\boldsymbol{\psi}_1}), \qquad c_{\boldsymbol{\psi}_1}(x, x^{'}) = \sigma_1^2(1 + \frac{|x - x^{'}|}{l_1} + \frac{|x - x^{'}|^2}{3l_1^2}) \exp(-\frac{|x - x^{'}|}{l_1}), \quad (6.14)$$

$$z_{2,\boldsymbol{\theta}} \sim GP(0, c_{\boldsymbol{\psi}_2}), \qquad c_{\boldsymbol{\psi}_2}(x, x^{'}) = \sigma_2^2(1 + \frac{|x - x^{'}|}{l_2} + \frac{|x - x^{'}|^2}{3l_2^2}) \exp(-\frac{|x - x^{'}|}{l_2}), \quad (6.15)$$

so we have $\boldsymbol{\psi}_1 = (\sigma_1, \sigma_{1,\text{mes}}, l_1)$ and $\boldsymbol{\psi}_2 = (\sigma_2, \sigma_{2,\text{mes}}, l_2)$.

The prior distributions for each variable are shown in table 6.4. Uniform priors are taken for the parameters as we do not favor their nominal values. For the variance hyperparameters, we take Jeffrey's priors to favor low-variance discrepancies that correspond to model predictions close to the observations. The remaining priors are uniform in the absence of meaningful information. Note that the variables are also constrained to a bounded domain, so their priors are truncated. The interval bounds for the variance hyperparameters are determined by order of magnitude of each observable. The bounds for correlation lengths are chosen with respect to the inter-observation distance and the channel radius.

The respective likelihood functions write:

$$\log \mathrm{p}_1(\mathbf{y}_{\text{obs},1}|\boldsymbol{\theta}, \boldsymbol{\psi}_1) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{C}_{1,\boldsymbol{\psi}_1}) - \frac{1}{2}(\mathbf{y}_{\text{obs},1} - \mathbf{f}_{1,\boldsymbol{\theta}})^T \mathbf{C}_{1,\boldsymbol{\psi}_1}^{-1}(\mathbf{y}_{\text{obs},1} - \mathbf{f}_{1,\boldsymbol{\theta}}) \quad (6.16)$$

Variable	Prior	Support
θ_i, $i \leq 4$	$\mathrm{p}(\theta_i) \propto 1$	$[0.5, 2]$
θ_5	$\mathrm{p}(\theta_5) \propto 1$	$[0.9, 1.1]$
σ_1	$\mathrm{p}(\sigma_1) \propto 1/\sigma_1^2$	$[10^{-4}, 1]$
σ_2	$\mathrm{p}(\sigma_2) \propto 1/\sigma_2^2$	$[10^{-5}, 5 \times 10^{-3}]$
$\sigma_{1,\mathrm{mes}}$	$\mathrm{p}(\sigma_{1,\mathrm{mes}}) \propto 1$	$[10^{-4}, 10^{-2}]$
$\sigma_{2,\mathrm{mes}}$	$\mathrm{p}(\sigma_{2,\mathrm{mes}}) \propto 1$	$[10^{-8}, 3 \times 10^{-5}]$
l_i, $i \leq 2$	$\mathrm{p}(l_i) \propto 1$	$[10^{-4}, 10^{-1}]$

Table 6.4: Prior distributions for model parameters and hyperparameters.

for the void fraction observations, and

$$\log \mathrm{p}_2(\mathbf{y}_{\mathrm{obs},1}|\boldsymbol{\theta}, \boldsymbol{\psi}_2) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det(\mathbf{C}_{2,\boldsymbol{\psi}_2}) - \frac{1}{2}(\mathbf{y}_{\mathrm{obs},2} - \mathbf{f}_{2,\boldsymbol{\theta}})^T \mathbf{C}_{2,\boldsymbol{\psi}_2}^{-1}(\mathbf{y}_{\mathrm{obs},2} - \mathbf{f}_{2,\boldsymbol{\theta}}) \tag{6.17}$$

for the bubble diameter, where we have noted for $i = 1, 2$ the following: $\mathbf{f}_{i,\boldsymbol{\theta}} = (f_i(x_{\mathrm{obs},1}, \boldsymbol{\theta}), \ldots, f_i(x_{\mathrm{obs},n}, \boldsymbol{\theta}))^T$; $\mathbf{C}_{i,\boldsymbol{\psi}_i}$ the matrix with (j,k) coefficient $c_{i,\boldsymbol{\psi}_i}(x_{\mathrm{obs},j}, x_{\mathrm{obs},k}) + \sigma_{i,\mathrm{mes}}^2 \delta_{j=k}$.

We note the vector containing all observations $\mathbf{y}_{\mathrm{obs}} = (\mathbf{y}_{\mathrm{obs},1}^T, \mathbf{y}_{\mathrm{obs},2}^T)^T$ and the vector of all hyperparameters $\boldsymbol{\psi} = (\boldsymbol{\psi}_1^T, \boldsymbol{\psi}_2^T)^T$ so the joint likelihood function writes:

$$\log \mathrm{p}(\mathbf{y}_{\mathrm{obs}}|\boldsymbol{\theta}, \boldsymbol{\psi}) = \log \mathrm{p}_1(\mathbf{y}_{\mathrm{obs},1}|\boldsymbol{\theta}, \boldsymbol{\psi}_1) + \log \mathrm{p}_2(\mathbf{y}_{\mathrm{obs},2}|\boldsymbol{\theta}, \boldsymbol{\psi}_2) \tag{6.18}$$

The posterior distributions are sampled using the M-H algorithm described in section 4.1.2, where the specificities of running each calibration technique are discussed in the last paragraphs. For each example presented here, the chains are well-converged, which is verified by their correlation lengths (see section 4.1.4). In figure 6.10 are plotted the self-correlation curves that correspond to the calibration operated in section 6.5.5. The mixing length corresponds to the area under the curves on these plots. The chains for the Bayes method exhibits larger mixing length, so poorer mixing due to its increased dimension. As a consequence, it needs to be run longer. In this example, the mixing length for KOH is around 37, for FMP 46, and Bayes 285. Thus we take respective chain lengths of 10^6, 10^6 and 7×10^6 steps. From each chain, a sample of size 4000 is extracted (see section 4.1.2), and the posterior distribution is estimated from this sample. For the hyperparameter posterior, it is obtained naturally within Bayes. For KOH it is a single-point value, and for FMP we take the distribution formed by the optimal hyperparameters for each parameter of the sample. Finally, the corrected predictions of the computer model can be obtained with the predictive equation (3.8), and we will also inspect the predictions of the computer model alone, obtained following equation (3.17).

The computation time for each method is also compared on the example of section 6.5.5, and reported in table 6.5. It is seen that the cost of the KOH calibration is mainly due to the search for its best hyperparameters, which requires a numerical technique

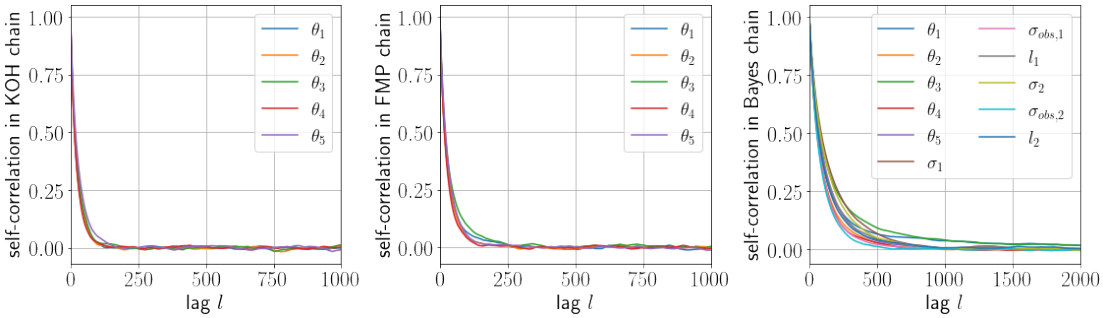


Figure 6.10: Self-correlation of the Markov chains for the KOH method (left), FMP (middle) and Bayes (right), in the study of section 6.5.5. Note that the x-axis on the Bayes plot does not have the same scale.

to integrate over the whole parameter domain. The FMP method was applied using two cycles of the adaptive sampling algorithm (section 4.3), with a total of 50+200+200 training points. The cost of the Bayes technique, as previously mentioned, is due to the bigger chain length required to keep the quality of mixing comparable to the other methods.

Method	**Steps**				**Total**
	Initial optimisation	Surrogate building	MCMC	Predictions	
KOH	500	0	105	36	641
FMP	0	738	125	40	903
Bayes	0	0	1052	36	1088

Table 6.5: Computation time (in seconds) for each calibration method in the study of section 6.5.5.

6.5.2 Building a surrogate model of NEPTUNE_CFD

The surrogate model for NEPTUNE_CFD is built by constructing Gaussian Processes (GPs) on the principal modes of the experimental data, as presented in section 4.2.1. The choice is made to build separate surrogates for the void fraction and the bubble diameter to perform the calibration on both observables separately, although the technique could be applied to the joint output.

We constructed a nDoE of 2040 NEPTUNE_CFD simulations, sampled using a QMC technique based on the low-discrepancy Halton sequence [Le Maitre and Knio, 2010]. This space-filling nDoE in the parameter space is represented on figure 6.11 (left). A second nDoE of size 2040 was also created using a LHS design, that will serve as a verification nDoE. It is represented on the right plot of figure 6.11. On each nDoE, a few simulation runs failed, so the effective size of the training nDoE is 2017 and the validation nDoE is 2025.

For the GPs of the surrogates, we use a mean zero and a product Matern 5/2 covari-

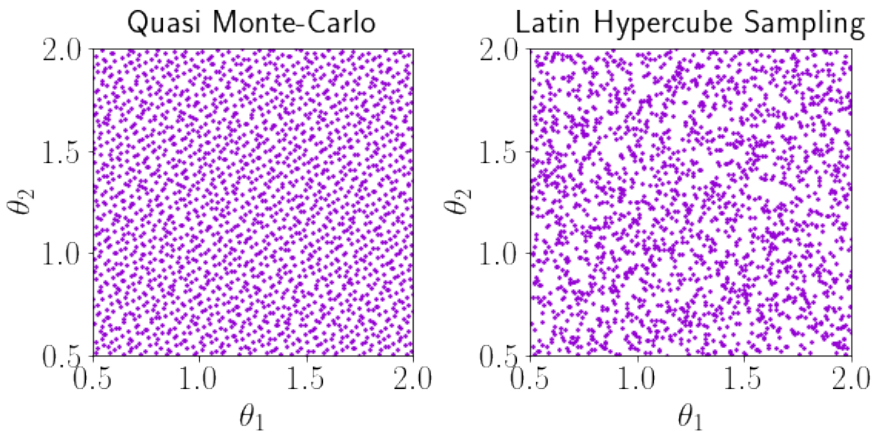


Figure 6.11: Projections of the five-dimensional nDoEs on two-dimensional planes. Left: training nDoE, right: validation nDoE.

ance function. The hyperparameters for these GPs are estimated by maximum likelihood. For each surrogate, the number of GPs used is equal to the number of principal modes selected. To determine the number d_* of modes, for each value of d_* we build the corresponding surrogate and plot its relative L2 error on the validation nDoE, given by equation (4.22).

In figure 6.12 are plotted the two types of error, along with the proportion of energy that is conserved, depending on d_* the number of modes selected. As the number of selected modes increases, the projection error is reduced, but the learning error increases since more GPs are needed. The optimal number of modes is determined when the projection error is approximately equal to the learning error and low overall error. Another selection criteria is that the projected energy is higher than 95% of the total energy, which is verified for $d_* > 2$ for the void fraction and $d_* > 2$ for the diameter. The surrogate for the void fraction provides a precision of 1% for $d_* = 3$, and the bubble diameter reaches around 2% of precision for $d_* = 6$. These values, represented on the figure as vertical red lines, are adopted in the rest of the calibration. From these figures, we consider that the surrogate model for NEPTUNE_CFD is sufficiently precise, and we use it as a substitute for f in the calibration equation without more consideration of the surrogate error.

6.5.3 First calibration without model error

We consider the case where no model error is assumed, and only the measurement uncertainty is present. Thus $z_{1,\boldsymbol{\theta}} = z_{2,\boldsymbol{\theta}} = 0$, so the likelihood function is a function of the squared residuals between the predictions and the observations. Consequently, parameters highlighted in the calibration will be the ones that minimize the fit to the data without any consideration of the shapes of the predictions.

The results are plotted on the top row of figure 6.13, where the model parameters underwent an affine transformation to the interval $[0, 1]$ for clarity. All observations are

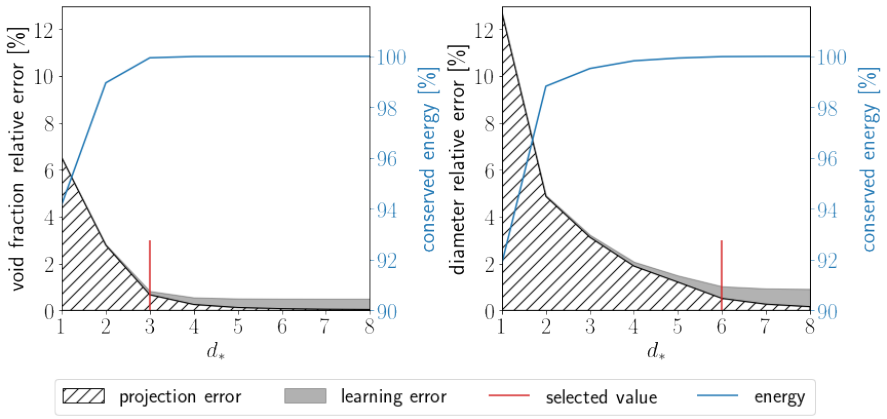


Figure 6.12: Error plots on the validation nDoE for the void fraction surrogate (left) and bubble diameter surrogate (right).

considered in this section, and we observe that the void fraction is fitted very well, but the diameter observations are not, with a very narrow confidence interval, so we find the effect of false confidence. Looking at the parameter posteriors on the right column, we see that these predictions correspond to very specific values of the parameters and that a slight modification of these coefficients would not be considered likely. Besides, they sometimes lie at the bounds of their range: both nucleation and breakup terms are set to -50%, for example. This reveals how a simple least-squares calibration can lead to misleading and not robust results, as we would not have strong confidence to use these calibrated parameters on other observables, seeing that they are overfitted to the void fraction.

To see what happens if the fit to the diameter observations was favoured, as an exercise, we add weight to this term in the likelihood function. The subsequent calibration results, shown in the bottom row of figure 6.13, now provide an excellent fit to the diameter, but the void fraction observations are globally overpredicted by 20% which is prohibitive. Note that the new coefficient values are completely different from the previous ones.

Two observations can be made from these results :

- There is no value of coefficients that provides an acceptable fit to void fraction and diameter observations simultaneously,
- it is thus necessary to use a non-zero model discrepancy term to obtain robust calibration results.

We now turn towards the calibration with the model discrepancy terms presented in 6.5.1.

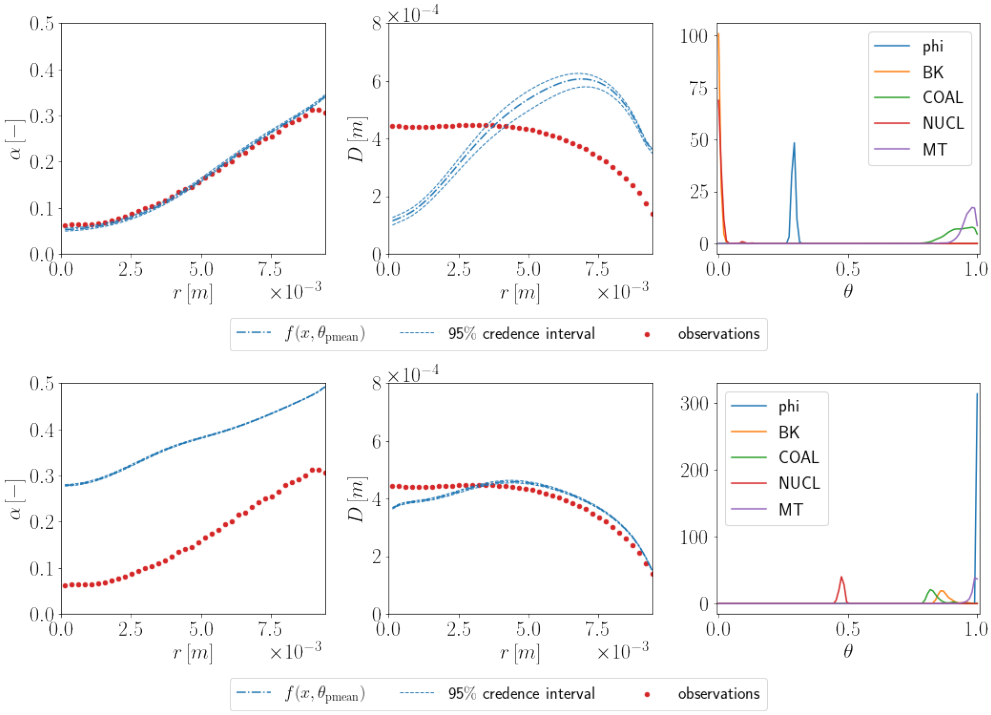


Figure 6.13: Top row, from left to right: predictions of void fraction, bubble diameter and parameter posteriors, with no model error. Bottom row: the equivalent figures when calibrating without model error, with a weighting favorable towards diameter observations.

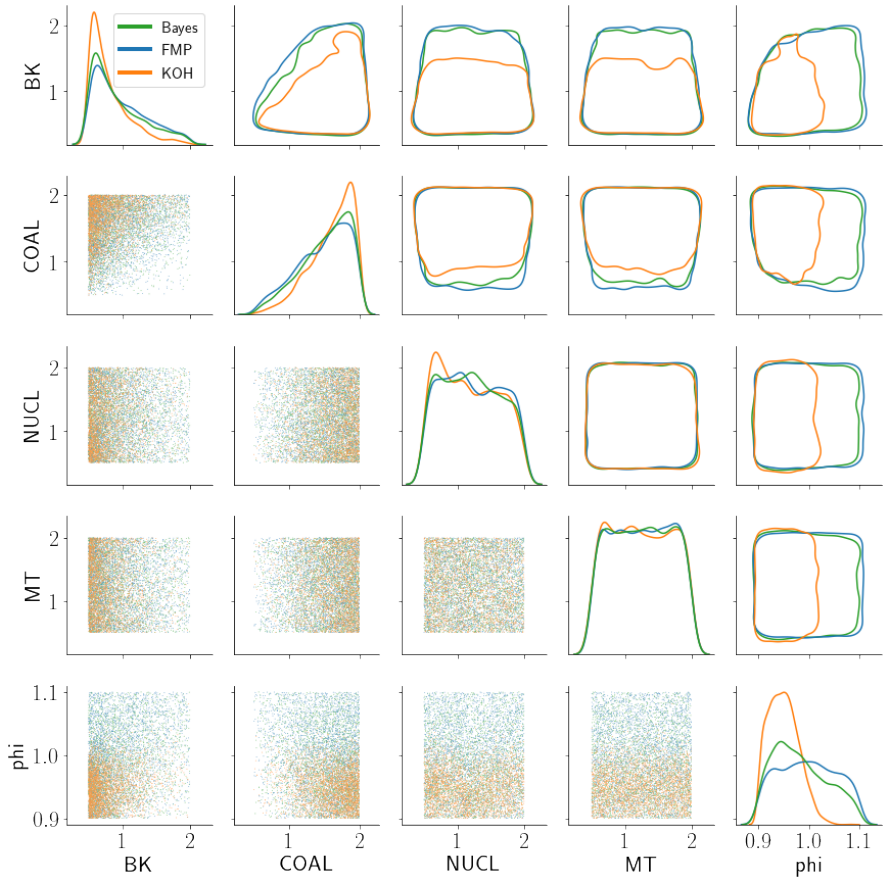


Figure 6.14: Posterior distribution of model parameters with the void fraction observations only. The plots on the upper diagonal represent 50% probability mass contours.

6.5.4 With model error, using void fraction observations

In this section, we infer the parameters of the NEPTUNE_CFD model using only the void fraction observations. The differences between the three calibration methods are examined, along with the updated predictions of the void fraction. We will also consider the predictions of bubble diameter that we call *blind predictions* because the corresponding observations are not yet included. In the next section, they will be included, and the impact on the calibration results will be examined.

Calibration results

The parameter posterior distributions are represented for each calibration technique on figure 6.14. To remind of their physical meaning we have noted $(\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$ as (BK,COAL,NUCL,MT,phi). The 1D marginals are represented on the diagonal, and the off-diagonal cases hold the 2D marginals. The three calibration techniques are differentiated by color. Since 1D and 2D marginals on the upper diagonal are estimated by KDE, a small portion of the probability mass is represented out of the support of parameters. In reality, all points of the sample are included in the support. The posterior distributions

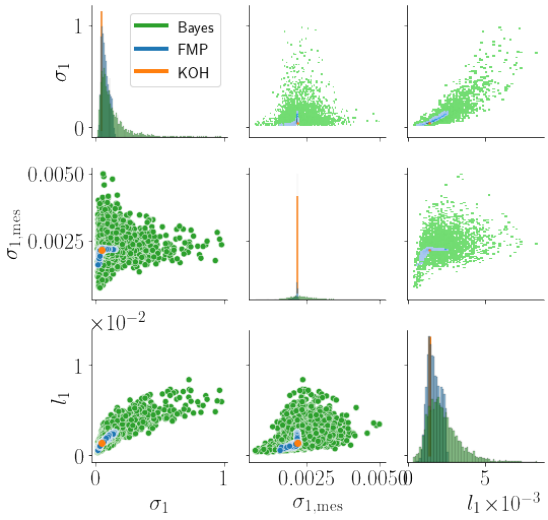


Figure 6.15: Posterior distribution of hyperparameters with the void fraction observations only.

obtained by the KOH method are slightly more peaky for the breakup and nucleation coefficients, and this is highlighted on the experimental flux distribution, where high values are deemed unlikely, whereas according to the Bayes and FMP calibration this is effect is not so pronounced.

The hyperparameter posteriors are plotted on figure 6.15. We remind that the KOH estimator is a single value; thus, it is represented as a vertical green line on the 1D marginals plots and a green point on the 2D marginals plots. Here we have $\widehat{\psi}_{1,\mathrm{KOH}} = (4.7 \times 10^{-2}, 2.2 \times 10^{-3}, 1.4 \times 10^{-3})$, so the model discrepancy is predominant as over 20 times the measurement error. The correlation length is approximately 1/7th of the size of the channel radius and around five times the distance between observations, so the model error is structured and well distinguished from the measurement error. There is also a high posterior correlation between σ_1 and l_1, due to the choice of covariance function, as we have seen in the numerical examples of previous chapters. The KOH estimator is well aligned with the maximum of the Bayes posterior, which corresponds to its definition. The Bayes posterior, however, presents a large variance that cannot be represented by a single value, which explains the more peaky behaviour of the parameter posteriors obtained by KOH calibration. The FMP estimation is also centered at the true most likely values, with a reduced variance as only optimal hyperparameters are considered. The additional variance helps to find the true parameter posteriors.

Predictions

The updated model predictions are represented on figure 6.16, by drawing samples of the statistical model for the truth $f + z_{\boldsymbol{\theta}}$, and with the model prediction at the posterior mean $f_{\boldsymbol{\theta}_{\mathrm{pmean}}}$ along with its associated 95% credible interval with size $3 * \sqrt{\mathrm{Var}_{\boldsymbol{\theta}}[f_{\boldsymbol{\theta}}]}$. The posterior samples of $f + z_{\boldsymbol{\theta}}$ are located at the observations and exhibit very low variance, which tells us that the estimated measurement error is very low and that

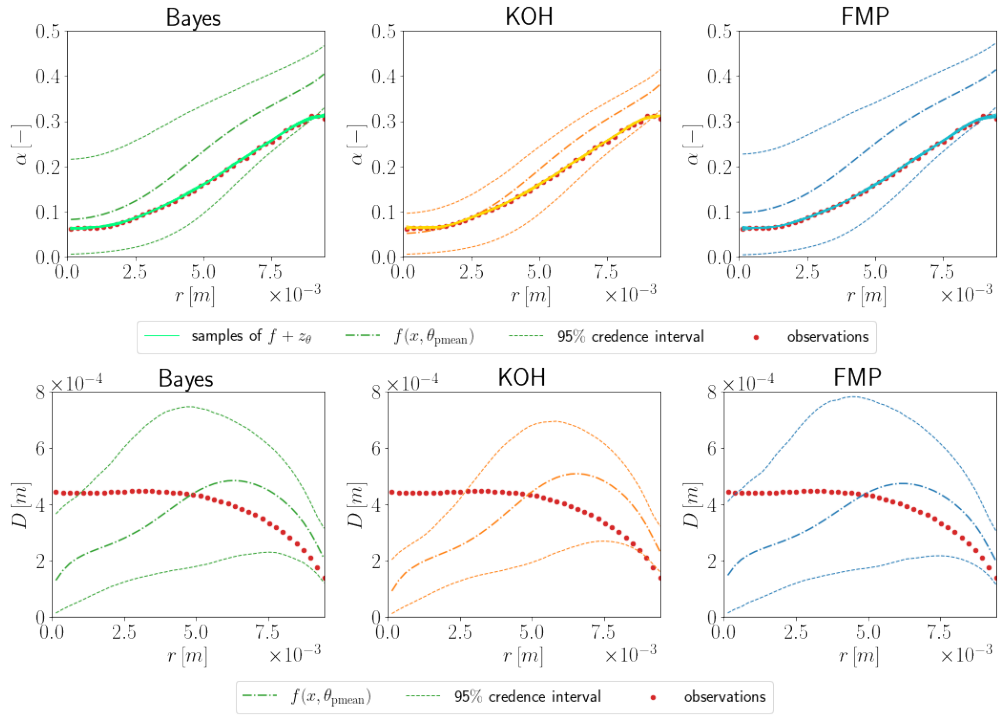


Figure 6.16: Predictions of void fraction (top row) and blind predictions of bubble diameter (bottom row), using void fraction observations only. Note that the legend is the same for all figures on the same row, with color depending on the calibration method.

the covariance function c_{ψ_1} provides a satisfying explanation of the observations. The most likely model-only prediction does not go through the observations, so a significant model error is inferred. The observations are still included in the 95% credible intervals. The FMP calibration provides the same quantitative results as the Bayes calibration, and the KOH method, due to its more peaky parameter posterior, show more narrow predictions.

The blind predictions on the bubble diameter, also represented in figure 6.16, show that the conservativism of the Bayes and FMP calibration allow to include the observations in the credible intervals, something which is not allowed in the KOH framework. Note also that we cannot draw samples of $f + z_{\boldsymbol{\theta}}$ here because the model discrepancy is a priori zero. The most likely prediction still shows the undesired plunging behaviour at the channel center.

Discussion

The physical interpretation of this study is that if we desire to improve the void fraction predictions, the breakup term should be lowered, the coalescence term increased, and the heat flux lowered. The nucleation and mass transfer terms do not influence this case. Note that the most likely predictions of void fraction do not differ much from the nominal calculation, which shows that there does not exist a "miracle value" of parameters that would significantly improve all predictions. From these results, we retain that some improvement of the physical models lies in the breakup and coalescence coefficients and that the hypothesis of a lower value of the experimental heat flux should be called into question.

In this example, the conservative advantage of the FMP calibration, compared to KOH, is put in the light. Due to the additional variance in hyperparameter estimation, the correct parameter posteriors were retrieved, and the credible intervals were accurately reproduced, which is preferred when making blind predictions on the bubble diameter.

6.5.5 Using void fraction and bubble diameter observations

We now take an interest on the calibration when the observations of both void fraction and bubble diameter are taken into account.

Calibration results

The parameter posteriors, shown in figure 6.17, tell a different story than the ones obtained with the void fraction only. The similarities are that the heat flux should still be lowered, and the nucleation term does not bear much influence. As a new find, we consider that the mass transfer term should be reduced, and, going against the results with void fraction only, the breakup should be increased and the coalescence reduced. There is also a noticeable correlation between these last two terms, which shows that they bear a similar effect on predictions. In this example, all three calibration methods provide similar results.

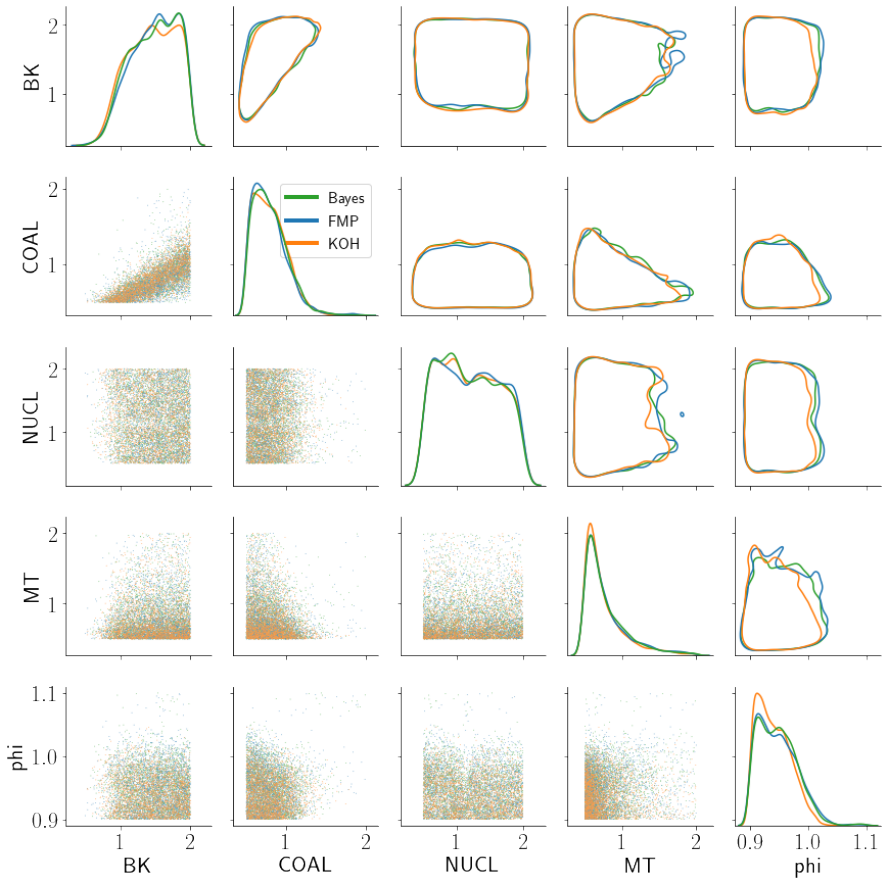


Figure 6.17: Posterior distribution of model parameters with the void fraction and bubble diameter observations. The plots on the upper diagonal represent 50% probability mass contours.

Figure 6.18: Posterior distribution of hyperparameters with the void fraction and bubble diameter observations.

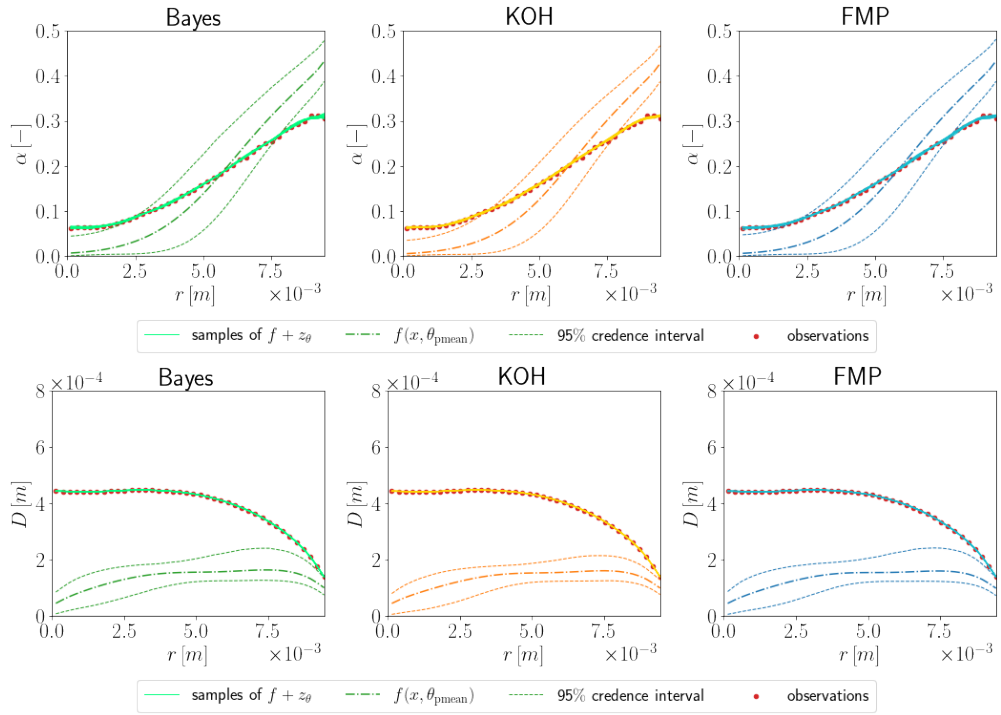


Figure 6.19: Predictions of void fraction (top row) and bubble diameter (bottom row), using void fraction and diameter observations. Note that the legend is the same for all figures on the same row, with color depending on the calibration method.

The posteriors for all six hyperparameters are plotted in figure 6.18. The KOH estimator is always represented with a vertical green line or a green point. This estimate corresponds to the maximum of the Bayes density, which confirms the numerical method used for its calculation. The FMP distributions are centered at the correct values, with reduced variance. From the posterior plots and the prediction plots that will follow, we see that this reduced variance still provides the correct conclusions. For each observable, there is a posterior correlation between model discrepancy intensity and correlation length, but no correlation is observed between both model discrepancies, which tends to confirm our prior assumption of independence of observations. The KOH estimator is $\widehat{\psi}_{\text{KOH}} = (9.8 \times 10^{-2}, 2.2 \times 10^{-3}, 1.8 \times 10^{-3}, 3.6 \times 10^{-4}, 7.4 \times 10^{-7}, 4.5 \times 10^{-3})$. As before, the intensity of both model discrepancy terms are predominant in from of the respective measurement errors, and the correlation lengths are approximately 1/4th and one half of the channel radius, so there is a noticeable structure in the model error.

Predictions

The predictions on both void fraction and bubble diameter are plotted in figure 6.19. It is interesting to note that the calibrated model predictions for the diameter present two characteristics: they are clearly below the observations, and they are quasi constant in the middle of the channel with a slight drop at the center. The 95% credible intervals with the model alone do not include any observation, which shows that a significant model

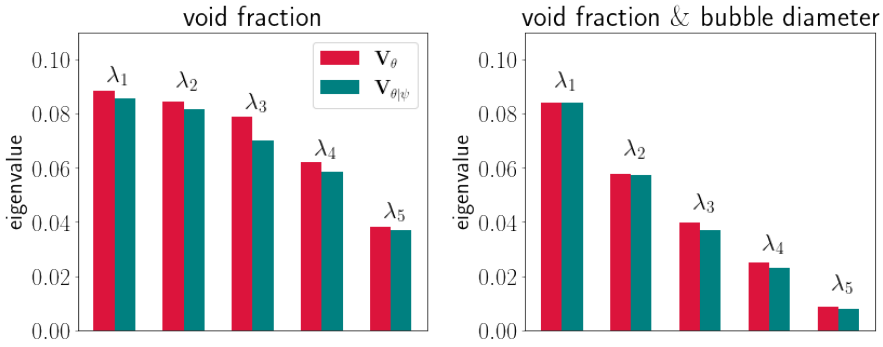


Figure 6.20: Ordered eigenvalues of the parameter posterior variance matrix (marginal and conditional), estimated from the Bayes sample, for both calibration studies. When the number of observations is higher (right figure), the posterior variance is smaller, corresponding to smaller eigenvalues with a faster decrease.

error is estimated. The samples of $f + z_{\boldsymbol{\theta},2}$ pass through the observations, so the data is correctly explained in the framework. The void fraction predictions with the model alone now show a different shape that predicts a low value at the center and a high value at the walls. The credible intervals include a significant part of the observations.

Posterior correlation between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$

In this paragraph, we try to understand why the KOH calibration performs almost exactly the same as the other two calibration techniques, whereas, in the example of the previous section, it showed a noticeably more peaky posterior distribution. In the absence of a universal method to determine a priori whether it is appropriate to apply the KOH method in a given situation, we give an interpretation inspired from the Laplace approximation study (section 3.3.1) by computing the posterior correlation between $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. It was proven in this section that the KOH method would be less accurate when this posterior correlation is high. This point is checked in our two calibration examples by using the posterior samples obtained in the Bayes method. From this, we estimate the parameter posterior variance $\mathbf{V}_{\boldsymbol{\theta}}$ and its conditional version $\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}$, which is theoretically the one that should correspond to the KOH posterior variance. The spectrum of both matrices is plotted in figure 6.20. This plot shows that in the calibration with void fraction only, the conditioning creates some reduction of the eigenvalues; thus, the KOH method provides an inaccurate estimation of the variance. The energy reduction due to the projection is obtained by $|\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}|/|\mathbf{V}_{\boldsymbol{\theta}|} \approx 1/1.7$. In the case with all observations, the projection has a smaller effect: $|\mathbf{V}_{\boldsymbol{\theta}|\boldsymbol{\psi}}|/|\mathbf{V}_{\boldsymbol{\theta}|} \approx 1/1.35$, so the KOH approximation is correct. This verification can only be done a posteriori, with the reference solution at hand.

Discussion

In this example, the three calibration techniques provide qualitatively the same results. The posterior distribution is unimodal with a low correlation between parameters and

hyperparameters, so the KOH approximation is appropriate in this case. The FMP calibration is shown to apply to a real problem with a computational cost reduced compared to the Bayes calibration, and the gain is expected to be more significant if more parameters or hyperparameters are considered.

The inclusion of diameter observations has noticeably changed the calibration procedure results. Since the diameter observations are predicted with great confidence by $f + z_{\boldsymbol{\theta},2}$, they carry significant weight in the results. We must answer why the calibration ends with model-only predictions that are arguably poor representations of the data. The reason is that we have chosen for both observables a covariance function that favors highly structured model discrepancies, in other words, flat functions $z_{1,\boldsymbol{\theta}}$ and $z_{2,\boldsymbol{\theta}}$. Naturally, if a value of model parameters predicted perfectly all observations, it would be found in the calibration results. Since this is not the case, the most likely model predictions are the ones that have an almost constant difference to the observations, in other words, that show *a similar shape to the observations*. Concretely, our choice of covariance function has favored predictions that limit the drop in diameter predictions at the channel center, which we had identified as undesired behaviour of the model. Thus, it is key to understand how the prior assumptions reflect on the results. We will see in section 6.5.6 the impact of different prior assumptions, including less structured model discrepancies. As a final remark, compare the void fraction predictions from figures 6.16 and 6.19: the first one is the one that fits the best the observations (since only the void fraction data was used), whereas in the second one a compromise is made with the diameter data. Yet the second predictions remain correct on average, which can be found by computing the section-averaged void fraction $\alpha_\mathrm{s} = 2\pi/S * \int_{0<r<r_\mathrm{max}} \alpha(r) r \mathrm{d}r$, where S is the section area. α_s is representative of the total quantity of gas in the test section. Experimentally, we have $\alpha_\mathrm{s} = 21\%$, with the predictions of void fraction only $\alpha_\mathrm{s} = 28\%$ and with all data $\alpha_\mathrm{s} = 23\%$. Thus, the same predictions can be either acceptable or undesirable from different perspectives. The calibration framework offers a natural way to encode the analyst's beliefs (should we emphasise predictions at the wall, average predictions in the section, avoid the drop in diameter at the center, etc.) to get the desired results from the data.

6.5.6 Sensitivity to the choice of model discrepancy kernel

In this section, we show that the choice of covariance function for $z_{\boldsymbol{\theta}}$ plays a determining role in calibration, so it must be carefully chosen to encode the analyst's true prior beliefs. We consider three additional kernels from the Matern family.

We explore the calibration results obtained with kernels Matern 1/2, Matern 3/2, and Squared Exponential. The first two are considered less structured than the Matern 5/2 because their correlation decreases faster at infinity as a function of the distance. The Squared Exponential shows the slightest decrease, so it is the most structured in the family. Note that the diagonal kernel, akin to measurement error, is less structured. These kernels were presented in section 2.3.3. We will show that using a less structured kernel tends to favor the fit to the data and more narrow posterior and predictions, whereas choosing a strong structure will emphasise the shape of the predictions and

provide larger confidence intervals.

The results for each calibration are plotted on figure 6.21, obtained with the FMP method. For all kernels, the samples of $f + z_{\boldsymbol{\theta}}$ go through all observations, so the real process is accurately represented. The predictions with Matern 1/2 resemble the ones without model error, albeit with larger posterior variance. From the Matern 3/2 onwards, the diameter is underpredicted so that the estimated void fraction is more acceptable. It is also interesting to notice that the posterior distribution for the wall heat flux changes according to the kernel, so that it is not clear whether its value should be heavily decreased, as there remains high posterior variance with the Squared Exponential case, with predictions virtually identical to the Matern 5/2 kernel.

6.6 Conclusion

In this chapter, we have presented a complete calibration study, where the successive steps were taken:

1. Description of the DEBORA configuration and assessment of the experimental uncertainties,
2. Study of the physical models in NEPTUNE_CFD and implementation of possible improvements,
3. Bayesian calibration following MCMC algorithms,
4. Physical analysis of the posterior results.

On the side of physics, we have demonstrated the influence of the wall heat flux on the void fraction predicted in the simulations, and by comparing it to the A6 data, we have found likely the possibility that the true heating value is below the assumed one. We also have proposed a modification of the interfacial area transport equation that impacts the bubble diameter predictions. We have limited the search of best parameters to regions close to the nominal values and found no configuration where the uncorrected model's predictions can correctly reproduce both void fraction and bubble diameter in the A6 case. Thus, there is a significant model error, as the difference between the data and the best-calibrated model predictions.

When comparing calibration techniques, we have found that the KOH approximation could provide confidence intervals smaller than the reference solution when the posterior correlation is present between parameters and hyperparameters, as was discussed previously in chapter 3. On the contrary, the FMP approximation allows to estimate posterior variance for hyperparameters, which, although smaller than the true value, provide an excellent approximation of the parameter posterior and calibrated predictions. The numerical cost of the FMP calibration stands in between KOH and the Bayes solution, and the computational gain is expected to be greater in problems with higher dimension since the cost of running a well-mixed MCMC on the total space increases significantly.

As we have already stressed, the calibration results depend strongly on how we have chosen to modify the NEPTUNE_CFD code and the statistical assumptions that com-

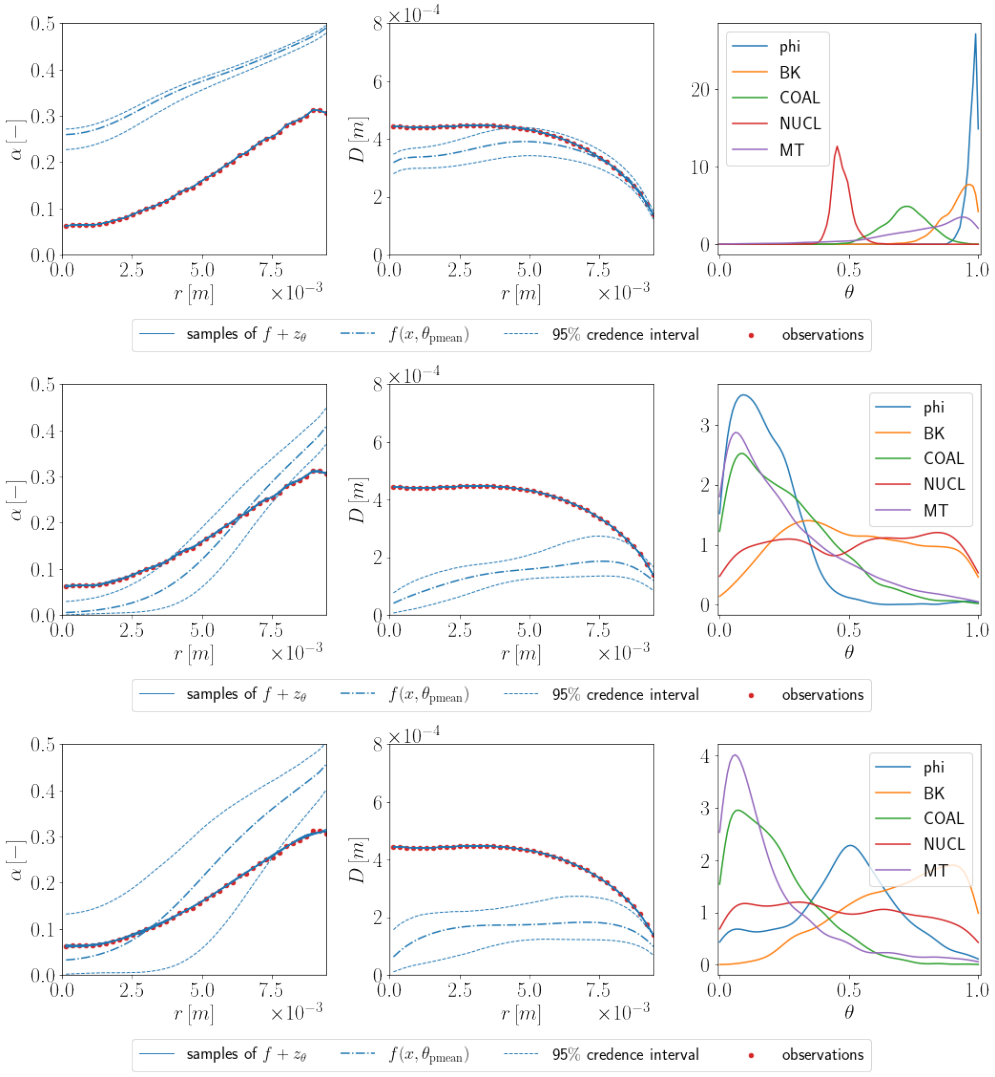


Figure 6.21: Left, middle and right columns: predictions of void fraction, bubble diameter and parameter posteriors. Top, middle and bottom rows: Matern 1/2, Matern 3/2, and Squared Exponential kernels.

prise the prior distributions and the choice of model discrepancy. A modification of the likelihood function is also possible to include specific physical insights, for example, to favor model parameters that predict the void fraction correctly at the wall. We hope to have given sufficient illustration about the practical interests of calibration. Although the act of translating physical insights into statistical assumptions in a Bayesian context is admittedly an arduous task, even for experts, the reward of getting accurate confidence intervals and uncertainty estimates is worth the effort.

Chapter 7

Conclusions and perspectives

7.1 Conclusions

This work was born from the need to accurately quantify the modelling uncertainties that are a central problem in CFD simulations, especially in the nuclear context where the experimental data is scarce and where there is an increasing need UQ methods for safety studies or TH simulations for design purposes. Fluid mechanics models generally involve many coefficients that are scenario-dependent, as well as experimental correlations that are established on large databases. Thus, the focus was on calibration techniques in the family of Bayesian inverse problems to determine model parameters from experimental data, taking into account experimental uncertainty, model error, and surrogate error.

In this work, we proposed a calibration technique that would fulfil criteria applicable in the nuclear context regarding the accuracy, robustness, and computational time. Taking a Bayesian point of view, we have started from a theoretical perspective by challenging the traditional definition of model error according to KOH. Then, we have proposed the FMP approximation and examined the validity of its underlying assumptions. The next step was to propose relevant algorithms to reduce the computational time of the method, which we have done by introducing the concept of hGP with an adaptive algorithm to place training points in the posterior plausible region. Finally, we have illustrated our approach on two concrete cases. In the first case, we have used a group of 13 experiments with the estimation of model error for each, while in the second one, we considered a highly non-linear model in industrial two-phase CFD code.

Our contributions are summarized as follows:

- A new definition for model error was proposed that challenges the notion of "true value" for the model parameters. After a thorough literature review about Bayesian calibration techniques, we assessed the hyperparameter estimation for the model discrepancy in the existing methods. Since calibration's goal is arguably to determine the parameter posteriors accurately, we proposed our definition to fulfil this objective better. The notion of "true value" was also identified as the source of identifiability problems between the model discrepancy and the parameter uncertainty, and those problems are solved in our formulation since, for each model prediction, a specific shape of model discrepancy is learned, leading to a more flexible estimation of the term that is relevant when the model predictions are sensitive to the parameters.

- A complete calibration framework was developed, in the vein of the KOH method, the Full Bayesian Analysis of Higdon, and the considerations of Plumlee. This framework allows to determine the plausible model parameters values in light of the data and make updated predictions, with model/discrepancy for predicting the true process, and model alone for evaluating its capabilities to recover the data, indicator of model error. The predictions also are provided with credence intervals, which give honest representations about the level of uncertainty that remain in both cases. Theoretical considerations helped us determine that our calibration framework was more accurate and robust than the KOH method with a study in the ideal case of a joint posterior in the form of a Gaussian mixture with well-separated modes. In a numerical study, we illustrated the problem of hyperparameter estimation in the non-microergodic case, which occurs with the commonly-used kernels in calibration problems. We determined that the calibration process would provide correct answers about the model parameters despite difficulties in estimating these quantities. In a second study, three asymptotic frameworks for kriging were presented. We showed that, in the calibration of both well-specified and misspecified models, the parameter posteriors exhibited normality. This argument favours calibration techniques since adding more observations will tend to increase the knowledge about parameters.

- The numerical execution of the calibration framework was carefully considered by exhibiting algorithms that are relevant to put it into practice. MCMC sampling of the posterior distribution was adopted, as it is common practice in this field, and we showed the subtility of its workings when applied to the three calibration frameworks studied in this thesis. We also exhibited an Importance Resampling algorithm that allowed us to refine the FMP sample to bring it closer to a sample of the true posterior, although due to the resulting thinning of the sample, it was not used in the applications. Surrogate modelling of the computer code using PCA coupled with Gaussian Processes (GPs) was proposed as a means to deal with high dimensional output space, which is typically the case for computer models. Then we proposed the hGP methodology to reduce the cost of the FMP calibration, which avoids dealing with an optimization problem at each step of the MCMC. Then, we proposed an adaptive algorithm for locating training points for the hGPs that would be plausible a posteriori. It is based on successive MCMCs with a resampling step of the points drawn, with weights given by the predictive variance of the hGPs, inducing a balance between exploration and exploitation. The approach in the algorithm is justified by showing that the approximation error of posterior distribution is controlled by the approximation error of the hGPs.

- A first application on a heat flux partitioning model in boiling was proposed. The aim of the study was twofold: first, see if the calibration process could improve the model predictions on the experiments; secondly, learn from the calibration the values of relevant parameters, namely the contact angle and two coefficients in the departure diameter correlation. This study has been done on a database of 13 experiments with a separate model discrepancy term for each, leading to 26 hyperparameters overall, which is a high value to treat in this kind of problem. After

considering the calibrations on each experiment individually, we have grouped the data in three categories relatively to how well the posterior predictions behaved. The reference solution could be obtained only using the nine well-behaved experiments, whereas the FMP method and a pooled version of KOH were performed on the totality. In all cases, we showed that the KOH method with separate estimations of hyperparameters failed when dealing with multiple experiments and that the FMP method was always closer to the reference solution.

- A second application was proposed, in which we dealt with a more complex computer model in the form of a 3D multiphase CFD solver. The initial step required careful consideration of the uncertainties by formulating physics-based assumptions to improve the model predictions on the DEBORA experiments. After testing the validity of these assumptions with multiple OAT analyses, and considering multiple levels of model modification, the calibration was setup by introducing multiplicative coefficients in front of the source terms of the interfacial area transport model, as well as on the wall heat flux which was identified as a possible source of experimental uncertainty. Two quantities of interest were considered, the void fraction and the bubble diameter. We operated a separate and joint treatment, showing how the robust character of the FMP method was favorable when predicting non-observed quantities. In all cases, the FMP method was closer to the reference solution and showed appropriate credence intervals on the predictions. Multiple kernels in the Matern family were tested to show how the choice of prior impacts the calibration results. Finally, we showed the importance of considering model error by proving how the calibration process is impaired when it is not included.

We find that calibration methods have proven effective in answering the questions targeted in this work, as they allow quantifying multiple sources of uncertainty and are an efficient way of considering the information (and lack of) obtained from experimental data. An honest representation of uncertainty is extremely valuable in scientific studies, and to ignore it willingly or not is a sure way to end up with false conclusions. In both our applications, we have dealt with consequent model error: despite our best attempts at improving the models' predictions by intrusive modifications, some experimental data remained out of the coverage of the credence intervals of the models. Thus two questions arise: What use is a computer model that we know is wrong? And what good are the conclusions about its parameters in this case?

In our opinion, answers to both questions are found in the database's width, the number and the variety of scenarios used for calibration if they are close to the new scenario for which we desire to use the model. In our application domain, we are unlikely to find some values of parameters for which the computer model will be perfect on all experiments. After calibration, it might provide satisfying predictions on, say, 90% of cases and unsatisfying on the other 10%. Thus, when considering a new scenario for prediction, we advise using parameter values that were determined on similar scenarios, as is already common practice when dealing with RANS turbulence models. Parameter values that are determined on a very large set of scenarios are useful when the new scenario is completely out of the known database, but do not offer a guarantee that

the predictions will be accurate. It might be the one time we are willing to bend the rules established by the *Bayesian desiderata*, which state that all sources of information should be considered, and none ignored. In practice, irrelevant information is discarded, like outliers. Of course, this must be done with caution, as sometimes, what we think to be an outlier is the hidden signal. In conclusion, we would instead use multiple models that work well in specific scenarios. When it is required of us to choose a single model to use in every case, the global performance of each candidate can be faithfully compared with calibration techniques.

7.2 Perspectives

We present a few ideas that can extend the present work. They are categorized into two groups: further developments on the methods and the next steps of the calibration studies presented as applications.

7.2.1 Extensions of the FMP framework

The two assumptions for the good working of the FMP approximation were given in section 3.1.3. Nonetheless, we show on the Gaussian mixture example that it tends to provide a biased estimation of the relative importance of posterior modes, which was observed in the applications. We note that the second assumption could be removed, stopping at the fourth equality in (3.13). This equality can be seen as a fixed-point equation for the true posterior, which could provide ideas of algorithms to get samples from a better approximation of the true posterior. This identity is used to correct the FMP sample in the Importance Resampling algorithm. It could be integrated as part of a MCMC procedure.

7.2.2 Improved methods for calibration in large dimension

As mentioned previously, probably the future of calibration methods lies in the inclusion of many experiments, i.e. many observations (high dimension of output space) and many parameters to calibrate in a simultaneous manner (high dimension of the input space). Thus, impactful applications could be realized if some techniques, already mature in machine learning, such as sparse GPs, would be used in a calibration framework.

Note the specific structure of the likelihood function: if one wished to take into account many experiments, considered to be independent of each other, with individual model discrepancies, the FMP method is perfectly adapted as optimal hyperparameters can be determined separately. Doing the same with KOH hyperparameters fails, as was shown in chapter 5: this method requires solving an optimization on the entire hyperparameter space, which is costly. The Bayes method requires MCMC sampling in high dimension also, which could be attempted with sampling methods that benefit from the quasi additive structure.

On another note, the Bayesian Committee Machine algorithm of [Tresp, 2000], in the case of clustered data, could also benefit from the idea of varying hyperparameters,

which would constitute an application to the Machine Learning field.

7.2.3 Considering new forms of model error

As we have discussed in section 2.3.3, knowledge about the model error can be introduced via the model discrepancy kernel, but in practice, it is difficult to formulate. Thus, it is essential to work with applications in which we possess this kind of knowledge and see how different kernels perform. Morally, kernels are a way to express similarities between two things: find a good measure to express what makes a model prediction physically acceptable and what does make it not acceptable, and the kernel can be derived from it.

In the same vein comes the prior mean of $z_{\boldsymbol{\theta}}$, which can be formulated from prior knowledge about the shape of the global model discrepancy. In most articles, it is advised to set it to zero - and in this thesis, this is what worked well for us. Still, it might be a good way to incorporate expert judgement.

One interrogation that goes against these ideas is the following: "If we have significant prior knowledge about the error, why not use it to improve the model instead of formulating a fancy model discrepancy ?". It is a good question, see its discussion in [Brynjarsdóttir and O'Hagan, 2014]. We answer that yes, practically it might be the best course to follow. Yet, you never know what you might find by looking into new directions.

7.2.4 Adaptive construction of surrogates for computer codes

We have already noted that most adaptive algorithms were used to build surrogates of the codes, not of the hyperparameters. An interesting perspective is thus to employ the ideas in our algorithm to the computer codes themselves. To the best of our knowledge, no such algorithm exists that would place training points in zones that are plausible a posteriori in the whole calibration framework, considering model error. One attempt in this direction is [Damblin, Barbillon, et al., 2018].

The present algorithm could be ameliorated as well, as it possesses user-determined parameters, such as the number of MCMCs, or the number of training points added in each step. What's more, the thinning of points caused by the resampling step could become costly, so other approaches such as a global minimization of the posterior-averaged prediction variance could prove fruitful.

7.2.5 Asymptotic frameworks in calibration

Theoretical results in asymptotic regimes in kriging are scarce and difficult to obtain. Most of them are established with a well-specified mean of the GP, which in the calibration problem is not the case due to the multiple model parameter values. Thus, it remains an open problem to establish such results exist in a calibration context. They would, however, have the benefit to provide some reassurance to the users of calibration

methods as a guarantee that the method remains consistent in the case of continuous data acquisition or data assimilation.

7.2.6 Data-informed modifications of the Ruyer-Seiler model

A frequent question that was asked to us when presenting our work to CFD practitioners is, "do the calibration techniques tell us how to improve the models?". Taking the question in its literal sense, the answer is no. They are only able to evaluate different models and compare them. But, they can do it systematically for a family of models. Thus, a physicist could formulate various experimental correlations, which could depend on different variables, or even be nested, and the calibration would do the rest. This work would constitute an impactful contribution to the calibration literature.

See [Nadiga et al., 2019] for an example of how calibration techniques aided a CFD practitioner to formulate a correlation in a turbulence model.

This type of investigations could be performed in the example of the Ruyer-Seiler model by trying more refined corrections, such as introducing space-dependent coefficients that could help determine for example if the breakup or coalescence terms are under- or over-estimated near the wall or in the channel center.

Appendix

Appendix A

Departure diameter data

In this Appendix, we list the experimental measurements of bubble departure diameter that were used in chapter 5 to evaluate the predictive ability of two experimental correlations: the one established by Kommajosyula, and the one obtained after calibrating the coefficients with the Kennel data.

Pressure p [bar]	Velocity V $[m.s^{-1}]$	Subcooling ΔT_{sub} [K]	Superheat ΔT_{sup} [K]	Diameter D_{d} [mm]
1.00	0.87	9.80	4.70	1.07
1.00	1.43	8.50	4.90	0.95
1.00	2.28	8.70	4.50	0.87
1.00	3.19	10.50	4.80	0.62
1.00	0.87	9.80	7.30	1.66
1.00	1.43	8.50	6.70	1.48
1.00	2.28	8.70	6.50	1.04
1.00	3.19	10.50	6.30	0.73
1.00	0.87	9.80	7.90	1.85
1.00	1.43	8.50	8.30	1.68
1.00	2.28	8.70	8.50	1.40
1.00	3.19	10.50	8.10	0.78

Table A.1: Departure diameter measurements from [Guan et al., 2016].

p [bar]	V [$m.s^{-1}$]	ΔT_{sub} [K]	ΔT_{sup} [K]	D_{d} [mm]
2.00	0.25	20.00	1.00	0.57
2.00	0.30	20.00	1.00	0.46
2.00	0.35	20.00	1.00	0.36
2.00	0.40	20.00	2.00	0.34
2.00	0.25	20.00	2.00	0.60
2.00	0.30	20.00	2.00	0.53
2.00	0.35	20.00	2.00	0.43
2.00	0.40	20.00	2.00	0.39
2.00	0.25	10.00	4.00	0.42
2.00	0.30	10.00	4.00	0.37
2.00	0.35	10.00	4.00	0.32
2.00	0.40	10.00	4.00	0.32
2.00	0.25	10.00	4.00	0.51
2.00	0.30	10.00	5.00	0.38
2.00	0.35	10.00	5.00	0.36
2.00	0.40	10.00	5.00	0.32
5.00	0.25	20.00	1.00	0.26
5.00	0.30	20.00	1.00	0.19
5.00	0.35	20.00	1.00	0.17
5.00	0.40	20.00	1.00	0.16
5.00	0.25	20.00	3.00	0.41
5.00	0.30	20.00	3.00	0.29
5.00	0.35	20.00	3.00	0.26
5.00	0.40	20.00	3.00	0.22

p [bar]	V [$m.s^{-1}$]	ΔT_{sub} [K]	ΔT_{sup} [K]	D_{d} [mm]
5.00	0.25	10.00	1.00	0.28
5.00	0.30	10.00	1.00	0.26
5.00	0.35	10.00	1.00	0.23
5.00	0.40	10.00	1.00	0.19
5.00	0.25	10.00	5.00	0.33
5.00	0.30	10.00	5.00	0.29
5.00	0.35	10.00	5.00	0.28
5.00	0.40	10.00	5.00	0.24
1.00	0.25	20.00	2.00	0.61
1.00	0.30	20.00	2.00	0.53
1.00	0.35	20.00	2.00	0.43
1.00	0.40	20.00	2.00	0.39
1.00	0.25	20.00	5.00	0.63
1.00	0.30	20.00	4.00	0.56
1.00	0.35	20.00	4.00	0.45
1.00	0.40	20.00	4.00	0.43
1.00	0.25	10.00	3.00	0.62
1.00	0.30	10.00	3.00	0.50
1.00	0.35	10.00	3.00	0.43
1.00	0.40	10.00	3.00	0.37
1.00	0.25	10.00	6.00	0.67
1.00	0.30	10.00	6.00	0.52
1.00	0.35	10.00	6.00	0.44
1.00	0.40	10.00	6.00	0.36

Table A.2: Departure diameter measurements from [Sugrue et al., 2014].

p [bar]	V $[m.s^{-1}]$	ΔT_{sub} [K]	ΔT_{sup} [K]	D_{d} [mm]
1.05	0.42	30.00	15.95	1.33
1.05	0.83	30.00	18.85	1.08
1.05	0.83	30.00	20.30	1.18
1.05	0.83	30.00	23.20	0.94
1.05	0.83	30.00	24.65	0.89
1.05	0.42	30.00	20.30	0.93
1.05	0.42	30.00	23.20	0.80
1.05	0.42	30.00	24.65	0.80
1.05	0.84	20.00	20.30	1.24
1.05	0.84	20.00	21.75	1.06
1.05	0.42	20.00	15.95	2.19
1.05	0.42	20.00	20.30	1.53
1.05	0.42	20.00	21.75	1.25
1.05	0.08	20.00	14.50	2.18
1.05	0.08	20.00	15.95	2.68
1.05	0.42	10.00	15.95	1.98
1.05	0.08	10.00	13.05	1.85
1.05	0.08	10.00	14.50	2.48
1.05	0.08	10.00	15.95	1.82
1.05	0.08	30.00	14.50	1.14
1.05	0.83	40.00	20.30	0.77
1.05	0.83	40.00	24.65	0.59
1.05	0.83	40.00	29.00	0.75
1.05	0.83	60.00	20.30	0.66
1.05	0.83	60.00	29.00	0.68
2.00	0.41	20.00	17.40	0.54
2.00	0.41	30.00	26.10	0.68
2.00	0.41	30.00	23.20	0.67
2.00	0.41	30.00	20.30	0.72
2.00	0.41	30.00	17.40	0.92
2.00	0.08	20.00	17.40	0.76
2.00	0.82	20.00	17.40	0.65
2.00	0.41	20.00	20.30	0.71
2.00	0.82	20.00	20.30	0.63
2.00	0.41	10.00	16.24	0.73
2.00	0.82	10.00	16.82	0.78
2.00	0.82	30.00	20.30	0.68
2.00	0.82	30.00	23.20	0.58
2.00	0.08	30.00	14.50	0.79
3.00	0.82	29.90	20.30	0.43
3.00	0.82	29.60	23.20	0.40
3.00	0.82	29.50	26.10	0.45
3.00	0.41	29.40	20.30	0.47
3.00	0.41	30.40	23.20	0.37
3.00	0.41	31.70	26.10	0.31
3.00	0.08	28.70	14.50	0.53
3.00	0.08	31.00	15.95	0.60
3.00	0.82	20.20	20.30	0.41
3.00	0.82	19.10	23.20	0.51
3.00	0.41	18.90	17.40	0.44
3.00	0.41	22.50	20.30	0.36
3.00	0.08	19.80	14.50	0.55
3.00	0.08	20.80	15.95	0.51
3.00	0.41	13.50	15.95	0.53

Table A.3: Departure diameter measurements from [Prodanovic et al., 2002]. The values for ΔT_{sup} were obtained from the analysis in [Kommajosyula, 2020].

Bibliography

AIAA (1998). *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*. American Institute of Aeronautics and Astronautics (cit. on pp. xxi, 7).

Aliprantis and Border (2006). *Infinite Dimensional Analysis: A Hitchhiker's Guide*. 3rd [rev. and enl.] ed. Berlin ; New York: Springer. 703 pp. ISBN: 978-3-540-29586-0 (cit. on p. 49).

Alvarez, Rosasco, and Lawrence (Apr. 16, 2012). "Kernels for Vector-Valued Functions: A Review". arXiv: 1106.6251 [cs, math, stat] (cit. on p. 82).

Andrieu and Thoms (Dec. 2008). "A Tutorial on Adaptive MCMC". In: *Statistics and Computing* 18.4, pp. 343–373. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-008-9110-y (cit. on pp. 74, 75).

Arendt, Apley, and Chen (Oct. 1, 2012). "Quantification of Model Uncertainty: Calibration, Model Discrepancy, and Identifiability". In: *Journal of Mechanical Design* 134.10, p. 100908. ISSN: 1050-0472, 1528-9001. DOI: 10.1115/1.4007390 (cit. on pp. xxv, 22).

Arendt, Apley, Chen, et al. (Oct. 1, 2012). "Improving Identifiability in Model Calibration Using Multiple Responses". In: *Journal of Mechanical Design* 134.10. ISSN: 1050-0472. DOI: 10.1115/1.4007573 (cit. on pp. xxv, 22).

ASME (2009). *Verification & Validation in Computational Fluid Dynamics & Heat Transfer*. American Society of Mechanical Engineers (cit. on pp. xxi, 7).

Baccou et al. (Sept. 1, 2020). "SAPIUM: A Generic Framework for a Practical and Transparent Quantification of Thermal-Hydraulic Code Model Input Uncertainty". In: *Nuclear Science and Engineering* 194.8-9, pp. 721–736. ISSN: 0029-5639. DOI: 10.1080/00295639.2020.1759310 (cit. on pp. xxii, 7).

Bachoc (Oct. 2013a). "Cross Validation and Maximum Likelihood Estimations of Hyper-Parameters of Gaussian Processes with Model Misspecification". In: *Computational Statistics & Data Analysis* 66, pp. 55–69. ISSN: 01679473. DOI: 10.1016/j.csda.2013.03.016 (cit. on pp. 22, 28, 62).

— (Jan. 1, 2013b). "Estimation Paramétrique de La Fonction de Covariance Dans Le Modèle de Krigeage Par Processus Gaussiens : Application à La Quantification Des Incertitudes En Simulation Numérique". thesis. Paris 7 (cit. on p. 28).

— (Mar. 2014). "Asymptotic Analysis of the Role of Spatial Sampling for Covariance Parameter Estimation of Gaussian Processes". In: *Journal of Multivariate Analysis* 125, pp. 1–35. ISSN: 0047259X. DOI: 10.1016/j.jmva.2013.11.015 (cit. on p. xxvii).

Bachoc et al. (2014). "Calibration and Improved Prediction of Computer Models by Universal Kriging". In: *Nuclear Science and Engineering* 176.1, pp. 81–97. DOI: 10.13182/NSE12-55 (cit. on p. 28).

Basu, Warrier, and Dhir (Mar. 15, 2005). "Wall Heat Flux Partitioning During Subcooled Flow Boiling: Part 1—Model Development". In: *Journal of Heat Transfer* 127.2, pp. 131–140. ISSN: 0022-1481. DOI: 10.1115/1.1842784 (cit. on pp. 101, 102).

Bauer, van der Wilk, and Rasmussen (May 30, 2017). "Understanding Probabilistic Sparse Gaussian Process Approximations". arXiv: 1606.04820 [stat] (cit. on p. 91).

Bayarri et al. (Oct. 2007). "Computer Model Validation with Functional Output". In: *The Annals of Statistics* 35.5, pp. 1874–1906. ISSN: 0090-5364. DOI: 10.1214/009053607000000163. arXiv: 0711.3271 (cit. on p. 82).

Bayarri et al. (May 1, 2007). "A Framework for Validation of Computer Models". In: *Technometrics* 49.2, pp. 138–154. ISSN: 0040-1706. DOI: 10.1198/004017007000000092 (cit. on pp. 18, 27, 41).

Bayes and Price (1763). "An Essay towards Solving a Problem in the Doctrine of Chances." In: *Philosophical Transactions (1683-1775)* 53, pp. 370–418. ISSN: 0260-7085. JSTOR: 105741 (cit. on p. 3).

Berge (1963). *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. New York: Macmillan Co. 270 pp. (cit. on pp. 49, 50).

Bestion (Dec. 1, 1990). "The Physical Closure Laws in the CATHARE Code". In: *Nuclear Engineering and Design* 124.3, pp. 229–245. ISSN: 0029-5493. DOI: 10.1016/0029-5493(90)90294-8 (cit. on pp. xxi, 6).

Beven and Freer (Aug. 1, 2001). "Equifinality, Data Assimilation, and Uncertainty Estimation in Mechanistic Modelling of Complex Environmental Systems Using the GLUE Methodology". In: *Journal of Hydrology* 249.1, pp. 11–29. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(01)00421-8 (cit. on p. 14).

Box and Wilson (1951). "On the Experimental Attainment of Optimum Conditions". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.1, pp. 1–45. ISSN: 0035-9246. JSTOR: 2983966 (cit. on p. 14).

Boyack et al. (May 1990). "Quantifying Reactor Safety Margins Part 1: An Overview of the Code Scaling, Applicability, and Uncertainty Evaluation Methodology". In: *Nuclear Engineering and Design* 119.1, pp. 1–15. ISSN: 00295493. DOI: 10.1016/0029-5493(90)90071-5 (cit. on pp. xxii, 7).

Boyle et al. (2005). *Multiple Output Gaussian Process Regression* (cit. on p. 36).

Brynjarsdóttir and O'Hagan (Oct. 2014). "Learning about Physical Parameters: The Importance of Model Discrepancy". In: *Inverse Problems* 30.11, p. 114007. ISSN: 0266-5611. DOI: 10.1088/0266-5611/30/11/114007 (cit. on pp. xxiv, 22, 29, 36, 167).

Caner (2020). *Analyse de Sensibilité de Simulations d'écoulements Bouillants*. Internal report. CEA (cit. on pp. 134, 137, 139).

Capodieci et al. (Nov. 1, 1998). "Novel Methodology for Postexposure Bake Calibration and Optimization Based on Electrical Linewidth Measurement and Process Metamodeling". In: *Journal of Vacuum Science & Technology B: Microelectronics and Nanometer Structures Processing, Measurement, and Phenomena* 16.6, pp. 3752–3758. ISSN: 1071-1023. DOI: 10.1116/1.590402 (cit. on p. 14).

Carmassi et al. (Mar. 25, 2019). "Bayesian Calibration of a Numerical Code for Prediction". arXiv: 1801.01810 [stat] (cit. on p. 22).

Chen, Simpson, and Ying (2000). "Infill Asymptotics for a Stochastic Process Model with Measurement Error". In: *Statistica Sinica* 10.1, pp. 141–156. ISSN: 1017-0405. JSTOR: 24306709 (cit. on p. 62).

Cole (1960). "A Photographic Study of Pool Boiling in the Region of the Critical Heat Flux". In: *AIChE Journal* 6.4, pp. 533–538. ISSN: 1547-5905. DOI: 10.1002/aic.690060405 (cit. on p. 101).

Conrad et al. (Oct. 1, 2016). "Accelerating Asymptotically Exact MCMC for Computationally Intensive Models via Local Approximations". In: *Journal of the American Statistical Association* 111.516, pp. 1591–1607. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2015.1096787. arXiv: 1402.1694 (cit. on p. 88).

Conti and O'Hagan (Mar. 1, 2010). "Bayesian Emulation of Complex Multi-Output and Dynamic Computer Models". In: *Journal of Statistical Planning and Inference* 140.3, pp. 640–651. ISSN: 0378-3758. DOI: 10.1016/j.jspi.2009.08.006 (cit. on p. 82).

Cox, Park, and Singer (July 28, 2001). "A Statistical Method for Tuning a Computer Code to a Data Base". In: *Computational Statistics & Data Analysis* 37.1, pp. 77–92. ISSN: 0167-9473. DOI: 10.1016/S0167-9473(00)00057-8 (cit. on pp. 14, 15).

Cox (Jan. 1, 1946). "Probability, Frequency and Reasonable Expectation". In: *American Journal of Physics* 14.1, pp. 1–13. ISSN: 0002-9505. DOI: 10.1119/1.1990764 (cit. on p. 3).

Craig et al. (June 1, 2001). "Bayesian Forecasting for Complex Systems Using Computer Simulators". In: *Journal of the American Statistical Association* 96.454, pp. 717–729. ISSN: 0162-1459. DOI: 10.1198/016214501753168370 (cit. on p. 14).

CSNI (2016). *PREMIUM: A Benchmark on the Quantification of the Uncertainty of the Physical Models in the System Thermal-Hydraulic Codes - Methodologies and Data Review*. Nuclear Energy Agency (cit. on pp. xxii, 7).

Cubizolles (Jan. 1, 1996). "Etude Stéréologique de La Topologie Des Écoulements Diphasiques à Haute Pression". These de doctorat. Ecully, Ecole centrale de Lyon (cit. on pp. 134, 135).

Currin (1988). "A Bayesian Approach to the Design and Analysis of Computer Experiments". In: DOI: 10.2172/814584 (cit. on pp. 14, 15, 25).

D'Auria, Lanfredini, and Muellner (2012). "Scaling Analysis in Bepu Licensing of LWR". In: *Nuclear Engineering and Technology* 44.6, pp. 611–622. ISSN: 1738-5733 (cit. on pp. xxii, 7).

D'Auria, Camargo, and Mazzantini (July 1, 2012). "The Best Estimate Plus Uncertainty (BEPU) Approach in Licensing of Current Nuclear Reactors". In: *Nuclear Engineering and Design* 248, pp. 317–328. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2012.04.002 (cit. on pp. xxii, 7).

Damblin, Barbillon, et al. (Jan. 2018). "Adaptive Numerical Designs for the Calibration of Computer Codes". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.1, pp. 151–179. ISSN: 2166-2525. DOI: 10.1137/15M1033162 (cit. on pp. 26, 95, 167).

Damblin and Gaillard (Feb. 13, 2019). "Bayesian Inference and Non-Linear Extensions of the CIRCE Method for Quantifying the Uncertainty of Closure Relationships Integrated into Thermal-Hydraulic System Codes". arXiv: 1902.04931 [stat] (cit. on pp. 7, 88).

Dashti and Stuart (2017). "The Bayesian Approach to Inverse Problems". In: *Handbook of Uncertainty Quantification*. Ed. by Ghanem, Higdon, and Owhadi. Cham: Springer International Publishing, pp. 311–428. ISBN: 978-3-319-12385-1. DOI: 10.1007/978-3-319-12385-1_7 (cit. on p. 5).

De Crecy and Bazin (2004). "Quantification of the Uncertainties of the Physical Models of CATHARE 2". In: BE-2004: International Meeting on Updates in Best Estimate Methods in Nuclear Installation Safety Analysis Proceedings. International Atomic Energy Agency (IAEA), p. 376 (cit. on pp. xxii, 7).

DeVolder et al. (Mar. 1, 2002). "Uncertainty Quantification for Multiscale Simulations". In: *Journal of Fluids Engineering* 124.1, pp. 29–41. ISSN: 0098-2202, 1528-901X. DOI: 10.1115/1.1445139 (cit. on p. 70).

Du, Zhang, and Mandrekar (2009). "Fixed-Domain Asymptotic Propoerties of Tapered Maximum Likelihood Estimators". In: *The Annals of Statistics* 37 (6A), pp. 3330–3361. ISSN: 0090-5364. JSTOR: 25662196 (cit. on p. 62).

Dubrule (Dec. 1, 1983). "Cross Validation of Kriging in a Unique Neighborhood". In: *Journal of the International Association for Mathematical Geology* 15.6, pp. 687–699. ISSN: 1573-8868. DOI: 10.1007/BF01033232 (cit. on p. 28).

Duvenaud, Lloyd, et al. (May 13, 2013). "Structure Discovery in Nonparametric Regression through Compositional Kernel Search". arXiv: 1302.4922 [cs, stat] (cit. on p. 36).

Duvenaud, Nickisch, and Rasmussen (Dec. 19, 2011). "Additive Gaussian Processes". arXiv: 1112.4394 [cs, stat] (cit. on p. 36).

EDF (2019). *Neptune_CFD Version 6.0.0 Verification and Validation*. Internal report. EDF (cit. on p. 135).

— (2020). *Neptune_CFD Version 6.0.0 Theory Guide*. Internal report. EDF (cit. on p. 130).

Estrada-Perez et al. (June 24, 2009). "PTV Experiments of Subcooled Boiling Flow Through a Rectangular Channel". In: 16th International Conference on Nuclear Engineering. American Society of Mechanical Engineers Digital Collection, pp. 929–939. DOI: 10.1115/ICONE16-48735 (cit. on p. 129).

Fisher (1925). "Statistical Methods for Research Workers". In: *Breakthroughs in Statistics: Methodology and Distribution*. Ed. by Kotz and Johnson. Springer Series in Statistics. New York, NY: Springer, pp. 66–70. ISBN: 978-1-4612-4380-9. DOI: 10.1007/978-1-4612-4380-9_6 (cit. on pp. 3, 14).

Flegal and Jones (Apr. 2010). "Batch Means and Spectral Variance Estimators in Markov Chain Monte Carlo". In: *The Annals of Statistics* 38.2, pp. 1034–1070. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/09-AOS735 (cit. on p. 79).

Fricker et al. (2011). "Probabilistic Uncertainty Analysis of an FRF of a Structure Using a Gaussian Process Emulator". In: *Mechanical Systems and Signal Processing* 25.8 (8), pp. 2962–2975. ISSN: 0888-3270 (cit. on p. 82).

Fuhg, Fau, and Nackenhorst (Aug. 18, 2020). "State-of-the-Art and Comparative Review of Adaptive Sampling Methods for Kriging". In: *Archives of Computational Methods in Engineering*. ISSN: 1134-3060, 1886-1784. DOI: 10.1007/s11831-020-09474-6 (cit. on p. 26).

Gamerman and Lopes (2006). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. 2nd ed. Texts in Statistical Science Series 68. Boca Raton: Taylor & Francis. 323 pp. ISBN: 978-1-58488-587-0 (cit. on p. 75).

Gardner et al. (Feb. 24, 2018). "Product Kernel Interpolation for Scalable Gaussian Processes". arXiv: 1802.08903 [cs, stat] (cit. on p. 34).

Garnier, Manon, and Cubizolles (2001). "Local Measurements on Flow Boiling of Refrigerant 12 in a Vertical Tube". In: *Multiphase Science and Technology* 13 (1&2). ISSN: 0276-1459, 1943-6181. DOI: 10.1615/MultScienTechn.v13.i1-2.10 (cit. on pp. 129, 133–135).

Garud, Karimi, and Kraft (Nov. 2017). "Design of Computer Experiments: A Review". In: *Computers & Chemical Engineering* 106, pp. 71–95. ISSN: 00981354. DOI: 10.1016/j.compchemeng.2017.05.010 (cit. on p. 25).

Gelman (2014). *Bayesian Data Analysis*. Third edition. Chapman & Hall/CRC Texts in Statistical Science. Boca Raton: CRC Press. 661 pp. ISBN: 978-1-4398-4095-5 (cit. on pp. 61, 63, 73, 74, 77, 117).

Gelman and Rubin (1992). "A Single Series from the Gibbs Sampler Provides a False Sense of Security". In: *Bayesian Statistics 4*. Oxford University Press, pp. 625–31 (cit. on p. 79).

Genton (2001). "Classes of Kernels for Machine Learning: A Statistics Perspective". In: p. 14 (cit. on p. 36).

Geweke (Dec. 1991). *Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments*. preprint. Staff Report. DOI: 10.21034/sr.148 (cit. on p. 79).

Gilman and Baglietto (Oct. 1, 2017). "A Self-Consistent, Physics-Based Boiling Heat Transfer Modeling Framework for Use in Computational Fluid Dynamics". In: *International Journal of Multiphase Flow* 95, pp. 35–53. ISSN: 0301-9322. DOI: 10.1016/j.ijmultiphaseflow.2017.04.018 (cit. on p. 100).

Glaeser (Mar. 26, 2008). "GRS Method for Uncertainty and Sensitivity Evaluation of Code Results and Applications". In: *Science and Technology of Nuclear Installations* 2008, e798901. ISSN: 1687-6075. DOI: 10.1155/2008/798901 (cit. on pp. xxii, 7).

Goldstein and O'Hagan (1996). "Bayes Linear Sufficiency and Systems of Expert Posterior Assessments". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 58.2, pp. 301–316. ISSN: 0035-9246. JSTOR: 2345978 (cit. on p. 14).

Gönen and Alpaydin (2011). "Multiple Kernel Learning Algorithms". In: *Journal of Machine Learning Research* 12.64, pp. 2211–2268. ISSN: 1533-7928 (cit. on p. 36).

Gregory and Smith (Jan. 1, 1990). "Calibration as Estimation". In: *Econometric Reviews* 9.1, pp. 57–89. ISSN: 0747-4938. DOI: 10.1080/07474939008800178 (cit. on p. 14).

Gu and Wang (Jan. 1, 2018). "Scaled Gaussian Stochastic Process for Computer Model Calibration and Prediction". In: *SIAM/ASA Journal on Uncertainty Quantification* 6.4, pp. 1555–1583. DOI: 10.1137/17M1159890 (cit. on p. 23).

Guan et al. (Jan. 2, 2016). "Effect of Bubble Contact Diameter on Bubble Departure Size in Flow Boiling". In: *Experimental Heat Transfer* 29.1, pp. 37–52. ISSN: 0891-6152. DOI: 10.1080/08916152.2014.926433 (cit. on pp. 102, 171).

Gueguen (Dec. 19, 2013). "Contribution à La Modélisation Multidimensionnelle Des Écoulements Bouillants Convectifs En Conduite Haute Pression Pour l'application Au Cas Des Réacteurs à Eau Pressurisée". These de doctorat. Grenoble (cit. on pp. 134–136).

Guelfi et al. (July 1, 2007). "NEPTUNE: A New Software Platform for Advanced Nuclear Thermal Hydraulics". In: *Nuclear Science and Engineering* 156.3, pp. 281–324. ISSN: 0029-5639. DOI: 10.13182/NSE05-98 (cit. on pp. 130, 132).

Gustafson (May 2005). "On Model Expansion, Model Contraction, Identifiability and Prior Information: Two Illustrative Scenarios Involving Mismeasured Variables". In: *Statistical Science* 20.2, pp. 111–140. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/088342305000000098 (cit. on pp. xxiv, 23).

Ha, Yun, and Jeong (Aug. 1, 2020). "Improvement of the Subcooled Boiling Model for Thermal–Hydraulic System Codes". In: *Nuclear Engineering and Design* 364, p. 110641. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2020.110641 (cit. on pp. 128, 129).

Han, Santner, and Rawlinson (Nov. 1, 2009). "Simultaneous Determination of Tuning and Calibration Parameters for Computer Experiments". In: *Technometrics* 51.4, pp. 464–474. ISSN: 0040-1706. DOI: 10.1198/TECH.2009.08126. pmid: 20523754 (cit. on p. 23).

Hibiki and Ishii (July 1, 2003). "Active Nucleation Site Density in Boiling Systems". In: *International Journal of Heat and Mass Transfer* 46.14, pp. 2587–2601. ISSN: 0017-9310. DOI: 10.1016/S0017-9310(03)00031-0 (cit. on p. 102).

Higdon, Gattiker, et al. (2008). "Computer Model Calibration Using High-Dimensional Output". In: *Journal of the American Statistical Association* 103.482, pp. 570–583. ISSN: 0162-1459. JSTOR: 27640080 (cit. on pp. xxv, 41, 70, 83).

Higdon, Kennedy, et al. (Jan. 1, 2004). "Combining Field Data and Computer Simulations for Calibration and Prediction". In: *SIAM Journal on Scientific Computing* 26.2, pp. 448–466. ISSN: 1064-8275. DOI: 10.1137/S1064827503426693 (cit. on pp. xxv, 29, 41, 70).

Hong et al. (Oct. 1, 2017). "Multi-Model Multivariate Gaussian Process Modelling with Correlated Noises". In: *Journal of Process Control* 58, pp. 11–22. ISSN: 0959-1524. DOI: 10.1016/j.jprocont.2017.08.004 (cit. on p. 36).

Hsu (Aug. 1, 1962). "On the Size Range of Active Nucleation Cavities on a Heating Surface". In: *Journal of Heat Transfer* 84.3, pp. 207–213. ISSN: 0022-1481. DOI: 10.1115/1.3684339 (cit. on p. 131).

INOUE et al. (July 1, 1995). "Void Fraction Distribution in BWR Fuel Assembly and Evaluation of Subchannel Code". In: *Journal of Nuclear Science and Technology* 32.7, pp. 629–640. ISSN: 0022-3131. DOI: 10.1080/18811248.1995.9731754 (cit. on p. 129).

Ishii and Hibiki (2006). *Thermo-Fluid Dynamics of Two-Phase Flow*. New York, N.Y: Springer Science+Business Media. 462 pp. ISBN: 978-0-387-28321-0 978-0-387-29187-1 (cit. on p. 129).

Ivanov et al. (May 13, 2018). "Uncertainty Analysis in Modelling for Light Water ReactorsConsistent Approach for Multi-scale Modelling". In: BEPU 2018 (cit. on pp. xxii, 7).

Jaynes and Bretthorst (2003). *Probability Theory: The Logic of Science*. Cambridge, UK; New York, NY: Cambridge University Press. ISBN: 978-0-511-06589-7 978-0-521-59271-0 978-0-511-06802-7 978-0-511-79042-3 978-1-139-63632-2 978-1-280-41722-1 (cit. on p. 4).

Jens and Lottes (May 1, 1951). *Analysis of Heat Transfer, Burnout, Pressure Drop and Density Data for High Pressure Water*. ANL-4627. Argonne National Lab. DOI: 10.2172/4421630 (cit. on p. 129).

Johnson, Moore, and Ylvisaker (Oct. 1, 1990). "Minimax and Maximin Distance Designs". In: *Journal of Statistical Planning and Inference* 26.2, pp. 131–148. ISSN: 0378-3758. DOI: 10.1016/0378-3758(90)90122-B (cit. on p. 25).

Jones, Schonlau, and Welch (1998). "Efficient Global Optimization of Expensive Black-Box Functions". In: *Journal of Global Optimization* 13.4, pp. 455–492. ISSN: 09255001. DOI: 10.1023/A:1008306431147 (cit. on p. 26).

Joucla and Probst (2008). "DIPE: Determination of Input Parameters Uncertainties Methodology Applied to CATHARE V2.5_1". In: *Journal of Power and Energy Systems* 2.1, pp. 409–420. DOI: 10.1299/jpes.2.409 (cit. on pp. xxii, 7).

Kejzlar et al. (Aug. 22, 2019). "Bayesian Averaging of Computer Models with Domain Discrepancies: A Nuclear Physics Perspective". arXiv: 1904.04793 [nucl-th, physics:physics, stat] (cit. on p. 29).

Kennedy and O'Hagan (2001a). "Bayesian Calibration of Computer Models". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.3, pp. 425–464. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00294 (cit. on pp. xxiv, 10, 15, 21, 30, 40).

— (2001b). "Supplementary Details on Bayesian Calibration of Computer Models". Internal report. Internal report (cit. on pp. 28, 40).

Kennel (1949). "Local Boiling of Water and Superheating of High Pressure Steam in Annuli". Thesis. Massachusetts Institute of Technology (cit. on pp. xxix, 99, 103).

Kim and Youn (Nov. 1, 2019). "Identifiability-Based Model Decomposition for Hierarchical Calibration". In: *Structural and Multidisciplinary Optimization* 60.5, pp. 1801–1811. ISSN: 1615-1488. DOI: 10.1007/s00158-019-02405-5 (cit. on pp. xxv, 22).

Kledy (May 30, 2018). "Développement d'une méthode de mesure des champs de vitesse et de température liquide en écoulement diphasique bouillant en conditions réacteurs ou simulantes". PhD thesis. Université Grenoble Alpes (cit. on p. 134).

Kleijnen, van Beers, and van Nieuwenhuyse (Sept. 1, 2012). "Expected Improvement in Efficient Global Optimization through Bootstrapped Kriging". In: *Journal of Global Optimization* 54.1, pp. 59–73. ISSN: 1573-2916. DOI: 10.1007/s10898-011-9741-y (cit. on p. 26).

Kocamustafaogullari and Ishii (Feb. 1, 1995). "Foundation of the Interfacial Area Transport Equation and Its Closure Relations". In: *International Journal of Heat and Mass Transfer* 38.3, pp. 481–493. ISSN: 0017-9310. DOI: 10.1016/0017-9310(94)00183-V (cit. on p. 101).

Kommajosyula (2020). "Development and Assessment of a Physics-Based Model for Subcooled Flow Boiling with Application to CFD". Massachusetts Institute of Technology. 148 pp. (cit. on pp. xxix, 100, 101, 103, 129, 173).

Končar and Mavko (Oct. 27, 2008). "Simulation of Boiling Flow Experiments Close to CHF with the Neptune_CFD Code". In: *Science and Technology of Nuclear Installations* 2008, e732158. ISSN: 1687-6075. DOI: 10.1155/2008/732158 (cit. on p. 134).

Krepper and Rzehak (Sept. 1, 2011). "CFD for Subcooled Flow Boiling: Simulation of DEBORA Experiments". In: *Nuclear Engineering and Design*. Seventh European Commission Conference on Euratom Research and Training in Reactor Systems (Fission Safety 2009) 241.9, pp. 3851–3866. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2011.07.003 (cit. on p. 134).

Kurul and Podowski (1990). "Multidimensional Effects in Forced Convection Subcooled Boiling". In: DOI: 10.1615/IHTC9.40 (cit. on pp. 129, 131).

Laplace (1814). "Essai philosophique sur les probabilités". In: (cit. on p. 3).

Le Maitre and Knio (2010). *Spectral Methods for Uncertainty Quantification: With Applications to Computational Fluid Dynamics*. Scientific Computation. Dordrecht [the Netherlands] ; New York: Springer. 536 pp. ISBN: 978-90-481-3519-6 (cit. on pp. 25, 147).

Leatherman, Dean, and Santner (Sept. 2017). "Designing Combined Physical and Computer Experiments to Maximize Prediction Accuracy". In: *Computational Statistics & Data Analysis* 113, pp. 346–362. ISSN: 01679473. DOI: 10.1016/j.csda.2016.07.013 (cit. on p. 26).

Lee, Park, and Lee (2002). "Local Flow Characteristics of Subcooled Boiling Flow of Water in a Vertical Concentric Annulus". In: DOI: 10.1016/S0301-9322(02)00026-5 (cit. on p. 129).

Li, Peng, and Ramadass (Oct. 1, 2009). "Accurate and Efficient Processor Performance Prediction via Regression Tree Based Modeling". In: *Journal of Systems Architecture* 55.10, pp. 457–467. ISSN: 1383-7621. DOI: 10.1016/j.sysarc.2009.09.004 (cit. on p. 26).

Ling, Mullins, and Mahadevan (Nov. 1, 2014). "Selection of Model Discrepancy Priors in Bayesian Calibration". In: *Journal of Computational Physics* 276, pp. 665–680. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2014.08.005 (cit. on pp. xxv, 22).

Liu, Bayarri, and Berger (Mar. 2009). "Modularization in Bayesian Analysis, with Emphasis on Analysis of Computer Models". In: *Bayesian Analysis* 4.1, pp. 119–150. ISSN: 1936-0975, 1931-6690. DOI: 10.1214/09-BA404 (cit. on pp. xxv, 19, 22, 27, 28).

Liu et al. (June 2020). "Gaussian Process Regression With Automatic Relevance Determination Kernel for Calendar Aging Prediction of Lithium-Ion Batteries". In: *IEEE Transactions on Industrial Informatics* 16.6, pp. 3767–3777. ISSN: 1941-0050. DOI: 10.1109/TII.2019.2941747 (cit. on p. 33).

Liu, Olewski, and Véchot (May 1, 2015). "Modeling of a Cryogenic Liquid Pool Boiling by CFD Simulation". In: *Journal of Loss Prevention in the Process Industries* 35, pp. 125–134. ISSN: 0950-4230. DOI: 10.1016/j.jlp.2015.04.006 (cit. on p. 98).

Loeppky, Bingham, and Welch (2006). *Computer Model Calibration or Tuning in Practice*, p. 20 (cit. on p. 29).

Loh and Lam (2000). "Estimating Structured Correlation Matrices in Smooth Gaussian Random Field Models". In: *Scopus*. ISSN: 0090-5364 (cit. on p. 62).

Lucor and Maitre (2018). "Cardiovascular Modeling With Adapted Parametric Inference". In: *ESAIM: Proceedings and Surveys* 62, pp. 91–107. DOI: 10.1051/proc/201862091 (cit. on pp. 88, 91, 95).

MacKay (1996). "Hyperparameters: Optimize, or Integrate Out?" In: *Maximum Entropy and Bayesian Methods: Santa Barbara, California, U.S.A., 1993*. Ed. by Heidbreder. Fundamental Theories of Physics. Dordrecht: Springer Netherlands, pp. 43–59. ISBN: 978-94-015-8729-7. DOI: 10.1007/978-94-015-8729-7_2 (cit. on p. 27).

— (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge, UK ; New York: Cambridge University Press. 628 pp. ISBN: 978-0-521-64298-9 (cit. on pp. 5, 51, 74).

Mañes et al. (June 30, 2014). "Validation of NEPTUNE-CFD Two-Phase Flow Models Using Experimental Data". In: *Science and Technology of Nuclear Installations* 2014, e185950. ISSN: 1687-6075. DOI: 10.1155/2014/185950 (cit. on p. 131).

Manon (Jan. 1, 2000). "Contribution à l'analyse et à La Modélisation Locale Des Écoulements Bouillants Sous-Saturés Dans Les Conditions Des Réacteurs à Eau Sous Pression". These de doctorat. Châtenay-Malabry, Ecole centrale de Paris (cit. on p. 134).

Mardia and Marshall (1984). "Maximum Likelihood Estimation of Models for Residual Covariance in Spatial Regression". In: *Biometrika* 71.1, pp. 135–146. ISSN: 0006-3444. DOI: 10.2307/2336405. JSTOR: 2336405 (cit. on pp. 61, 62).

Marzouk and Xiu (Jan. 1, 2009). "A Stochastic Collocation Approach to Bayesian Inference in Inverse Problems". In: *PRISM: NNSA Center for Prediction of Reliability, Integrity and Survivability of Microsystems* (cit. on p. 89).

Matheron (Dec. 1, 1963). "Principles of Geostatistics". In: *Economic Geology* 58.8, pp. 1246–1266. ISSN: 0361-0128. DOI: 10.2113/gsecongeo.58.8.1246 (cit. on pp. 14, 24).

Maupin and Swiler (Jan. 21, 2020). "Model Discrepancy Calibration Across Experimental Settings". In: *Reliability Engineering & System Safety*, p. 106818. ISSN: 0951-8320. DOI: 10.1016/j.ress.2020.106818 (cit. on p. xxiv).

Mchutchon and Rasmussen (2011). "Gaussian Process Training with Input Noise". In: *Advances in Neural Information Processing Systems*. Vol. 24. Curran Associates, Inc. (cit. on pp. 108, 109).

Mckay, Beckman, and Conover (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code". In: *Technometrics* 42.1, pp. 55–61. ISSN: 0040-1706, 1537-2723. DOI: 10.1080/00401706.2000.10485979 (cit. on pp. 14, 25).

Micchelli and Pontil (Dec. 1, 2004). "Kernels for Multi-Task Learning". In: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. NIPS'04. Cambridge, MA, USA: MIT Press, pp. 921–928 (cit. on p. 82).

Miller and Frenklach (1983). "Sensitivity Analysis and Parameter Estimation in Dynamic Modeling of Chemical Kinetics". In: *International Journal of Chemical Kinetics* 15.7, pp. 677–696. ISSN: 1097-4601. DOI: 10.1002/kin.550150709 (cit. on p. 14).

Montgomery (2013). *Design and Analysis of Experiments*. Eighth edition. Hoboken, NJ: John Wiley & Sons, Inc. 730 pp. ISBN: 978-1-118-14692-7 (cit. on p. 25).

Morris, Mitchell, and Ylvisaker (Aug. 1, 1993). "Bayesian Design and Analysis of Computer Experiments: Use of Derivatives in Surface Prediction". In: *Technometrics* 35.3, pp. 243–255. ISSN: 0040-1706. DOI: 10.1080/00401706.1993.10485320 (cit. on p. 15).

Nadiga, Jiang, and Livescu (Oct. 1, 2019). "Leveraging Bayesian Analysis to Improve Accuracy of Approximate Models". In: *Journal of Computational Physics* 394, pp. 280–297. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2019.05.015 (cit. on p. 168).

Nagel, Rieckermann, and Sudret (Mar. 1, 2020). "Principal Component Analysis and Sparse Polynomial Chaos Expansions for Global Sensitivity Analysis and Model Calibration: Application to Urban Drainage Simulation". In: *Reliability Engineering & System Safety* 195, p. 106737. ISSN: 0951-8320. DOI: 10.1016/j.ress.2019.106737 (cit. on p. 83).

Nukiyama (Dec. 1, 1966). "The Maximum and Minimum Values of the Heat Q Transmitted from Metal to Boiling Water under Atmospheric Pressure". In: *International Journal of Heat and Mass Transfer* 9.12, pp. 1419–1433. ISSN: 0017-9310. DOI: 10.1016/0017-9310(66)90138-4 (cit. on p. 98).

Picheny et al. (July 1, 2010). "Adaptive Designs of Experiments for Accurate Approximation of a Target Region". In: *Journal of Mechanical Design* 132.7, p. 071008. ISSN: 1050-0472, 1528-9001. DOI: 10.1115/1.4001873 (cit. on p. 26).

Plumlee (July 3, 2017). "Bayesian Calibration of Inexact Computer Models". In: *Journal of the American Statistical Association* 112.519, pp. 1274–1285. ISSN: 0162-1459, 1537-274X. DOI: 10.1080/01621459.2016.1211016 (cit. on pp. xxv, 23, 41, 70).

— (2019). "Computer Model Calibration with Confidence and Consistency". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 81.3, pp. 519–545. ISSN: 1467-9868. DOI: 10.1111/rssb.12314 (cit. on p. 23).

Poole and Raftery (2000). "Inference for Deterministic Simulation Models: The Bayesian Melding Approach". In: *Journal of the American Statistical Association* 95.452, pp. 1244–1255. ISSN: 0162-1459. DOI: 10.2307/2669764. JSTOR: 2669764 (cit. on p. 15).

Pratola et al. (May 1, 2013). "Fast Sequential Computer Model Calibration of Large Nonstationary Spatial-Temporal Processes". In: *Technometrics* 55.2, pp. 232–242. ISSN: 0040-1706. DOI: 10.1080/00401706.2013.775897 (cit. on pp. 26, 88).

Prince and Blanch (1990). "Bubble Coalescence and Break-up in Air-Sparged Bubble Columns". In: *AIChE Journal* 36.10, pp. 1485–1499. ISSN: 1547-5905. DOI: 10.1002/aic.690361004 (cit. on p. 133).

Prodanovic, Fraser, and Salcudean (Jan. 1, 2002). "Bubble Behavior in Subcooled Flow Boiling of Water at Low Pressures and Low Flow Rates". In: *International Journal of Multiphase Flow* 28.1, pp. 1–19. ISSN: 0301-9322. DOI: 10.1016/S0301-9322(01)00058-1 (cit. on pp. 102, 173).

Pronzato (Mar. 2017). "Minimax and Maximin Space-Filling Designs: Some Properties and Methods for Construction". In: *Journal de la Societe Française de Statistique* 158.1, pp. 7–36 (cit. on p. 25).

Raftery, Givens, and Zeh (June 1, 1995). "Inference from a Deterministic Population Dynamics Model for Bowhead Whales". In: *Journal of the American Statistical Association* 90.430, pp. 402–416. ISSN: 0162-1459. DOI: 10.1080/01621459.1995.10476529 (cit. on p. 14).

Ranz and Marshall (1952). "Evaporation from Drops". In: *Chemical Engineering Progress* 48, pp. 141–146 (cit. on p. 133).

Rasmussen and Williams (2006). *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. Cambridge, Mass: MIT Press. 248 pp. ISBN: 978-0-262-18253-9 (cit. on pp. 20, 34, 36, 48, 53).

Richenderfer et al. (Dec. 1, 2018). "Investigation of Subcooled Flow Boiling and CHF Using High-Resolution Diagnostics". In: *Experimental Thermal and Fluid Science* 99, pp. 35–58. ISSN: 0894-1777. DOI: 10.1016/j.expthermflusci.2018.07.017 (cit. on p. 102).

Robert (2007). *The Bayesian Choice*. Springer Texts in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-71598-8. DOI: 10.1007/0-387-71599-1 (cit. on pp. 5, 29, 138).

Robert and Casella (2010). *Monte Carlo Statistical Methods*. 2. ed., softcover reprint of the hardcover 2. ed. 2004. Springer Texts in Statistics. New York, NY: Springer. 645 pp. ISBN: 978-1-4419-1939-7 978-1-4757-4145-2 (cit. on pp. 73, 75).

Roma et al. (Aug. 1, 2021). "A Bayesian Framework of Inverse Uncertainty Quantification with Principal Component Analysis and Kriging for the Reliability Analysis of Passive Safety Systems". In: *Nuclear Engineering and Design* 379, p. 111230. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2021.111230 (cit. on p. 83).

Roy, Velidandla, and Kalra (Nov. 1, 1997). "Velocity Field in Turbulent Subcooled Boiling Flow". In: *Journal of Heat Transfer* 119.4, pp. 754–766. ISSN: 0022-1481. DOI: 10.1115/1.2824180 (cit. on p. 129).

Roy (Oct. 16, 2019). "Convergence Diagnostics for Markov Chain Monte Carlo". arXiv: 1909.11827 [stat] (cit. on p. 79).

Ruyer and Seiler (Aug. 1, 2009). "Modélisation Avancée de La Polydispersion En Taille Des Écoulements Bouillants". In: *La Houille Blanche* 95.4, pp. 65–71. ISSN: 0018-6368, 1958-5551. DOI: 10.1051/lhb/2009046 (cit. on pp. 132–134).

Sacks et al. (Nov. 1989). "Design and Analysis of Computer Experiments". In: *Statistical Science* 4.4, pp. 409–423. ISSN: 0883-4237, 2168-8745. DOI: 10.1214/ss/1177012413 (cit. on pp. 14, 26).

Saha and Zuber (Jan. 1, 1974). *Point of Net Vapor Generation and Vapor Void Fraction in Subcooled Boiling*. Scripta Book Co., Washington, DC (cit. on p. 129).

Sebastiani and Wynn (2000). "Maximum Entropy Sampling and Optimal Bayesian Experimental Design". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 62.1, pp. 145–157. ISSN: 1467-9868. DOI: 10.1111/1467-9868.00225 (cit. on p. 25).

Shewry and Wynn (Jan. 1, 1987). "Maximum Entropy Sampling". In: *Journal of Applied Statistics* 14.2, pp. 165–170. ISSN: 0266-4763. DOI: 10.1080/02664768700000020 (cit. on p. 25).

Situ et al. (Nov. 1, 2004). "Flow Structure of Subcooled Boiling Flow in an Internally Heated Annulus". In: *International Journal of Heat and Mass Transfer* 47.24, pp. 5351–5364. ISSN: 0017-9310. DOI: 10.1016/j.ijheatmasstransfer.2004.06.035 (cit. on pp. 102, 129).

Song and Fan (Feb. 1, 2021). "Temperature Dependence of the Contact Angle of Water: A Review of Research Progress, Theoretical Understanding, and Implications for Boiling Heat Transfer". In: *Advances in Colloid and Interface Science* 288, p. 102339. ISSN: 0001-8686. DOI: 10.1016/j.cis.2020.102339 (cit. on p. 102).

Stein (1999). *Interpolation of Spatial Data*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-1-4612-7166-6 978-1-4612-1494-6. DOI: 10.1007/978-1-4612-1494-6 (cit. on pp. xxvii, 33, 62).

Stuart (May 2010). "Inverse Problems: A Bayesian Perspective". In: *Acta Numerica* 19, pp. 451–559. ISSN: 1474-0508, 0962-4929. DOI: 10.1017/S0962492910000061 (cit. on p. 5).

Sugrue, Buongiorno, and McKrell (Sept. 2014). "An Experimental Study of Bubble Departure Diameter in Subcooled Flow Boiling Including the Effects of Orientation Angle, Subcooling, Mass Flux, Heat Flux, and Pressure". In: *Prof. Buongiorno via Chris Sherratt*. ISSN: 0029-5493 (cit. on pp. 102, 172).

Sundararajan and Keerthi (May 1, 2001). "Predictive Approaches for Choosing Hyperparameters in Gaussian Processes". In: *Neural Computation* 13.5, pp. 1103–1118. ISSN: 0899-7667. DOI: 10.1162/08997660151134343 (cit. on p. 28).

Thorn et al. (1965). "Boiling in Sub-Cooled Water during Flow up Heated Tubes or Annuli". In: DOI: 10.1243/PIME_CONF_1965_180_117_02 (cit. on p. 129).

Tian and Zhou (May 15, 1992). "The Maximum Theorem and the Existence of Nash Equilibrium of (Generalized) Games without Lower Semicontinuities". In: *Journal of Mathematical Analysis and Applications* 166.2, pp. 351–364. ISSN: 0022-247X. DOI: 10.1016/0022-247X(92)90302-T (cit. on p. 50).

Tolubinsky and Kostanchuk (1970). "Vapour Bubbles Growth Rate and Heat Transfer Intensity at Subcooled Water Boiling". In: International Heat Transfer Conference 4. Begel House Inc. DOI: 10.1615/IHTC4.250 (cit. on p. 101).

Tresp (Nov. 2000). "A Bayesian Committee Machine". In: *Neural Computation* 12.11, pp. 2719–2741. ISSN: 0899-7667, 1530-888X. DOI: 10.1162/089976600300014908 (cit. on p. 166).

Tuo and Wu (Aug. 28, 2015a). "A Theoretical Framework for Calibration in Computer Models: Parametrization, Estimation and Convergence Properties". arXiv: 1508.07155 [stat] (cit. on pp. xxv, 22, 41, 70).

— (July 26, 2015b). "Efficient Calibration for Imperfect Computer Models". arXiv: 1507.07280 [stat] (cit. on pp. 22, 66, 68).

— (Mar. 3, 2017). "Prediction Based on the Kennedy-O'Hagan Calibration Model: Asymptotic Consistency and Other Properties". arXiv: 1703.01326 [math, stat] (cit. on p. 22).

Unal (1976). "Maximum Bubble Diameter, Maximum Bubble-Growth Time and Bubble-Growth Rate during the Subcooled Nucleate Flow Boiling of Water up to 177 MN/M2". In: *International Journal of Heat and Mass Transfer* 19.6, pp. 643–649 (cit. on p. 133).

Vehtari et al. (Feb. 23, 2021). "Pareto Smoothed Importance Sampling". arXiv: 1507.02646 [stat] (cit. on p. 78).

Wilkinson (2010). "Bayesian Calibration of Expensive Multivariate Computer Experiments". In: *Large-Scale Inverse Problems and Quantification of Uncertainty*. John Wiley & Sons, Ltd, pp. 195–215. ISBN: 978-0-470-68585-3. DOI: 10.1002/9780470685853.ch10 (cit. on p. 83).

Wilson and Izmailov (Feb. 20, 2020). "Bayesian Deep Learning and a Probabilistic Perspective of Generalization". arXiv: 2002.08791 [cs, stat] (cit. on p. 22).

Wu, Wang, and Xiao (Oct. 2016). "A Bayesian Calibration-Prediction Method for Reducing Model-Form Uncertainties with Application in RANS Simulations". In: *Flow, Turbulence and Combustion* 97.3, pp. 761–786. ISSN: 1386-6184, 1573-1987. DOI: 10.1007/s10494-016-9725-6. arXiv: 1510.06040 (cit. on p. 37).

Wu et al. (Apr. 1, 1998). "One-Group Interfacial Area Transport in Vertical Bubbly Flow". In: *International Journal of Heat and Mass Transfer* 41.8, pp. 1103–1112. ISSN: 0017-9310. DOI: 10.1016/S0017-9310(97)00167-1 (cit. on p. 133).

Wu, Shirvan, and Kozlowski (Nov. 1, 2019). "Demonstration of the Relationship between Sensitivity and Identifiability for Inverse Uncertainty Quantification". In: *Journal of Computational Physics* 396, pp. 12–30. ISSN: 0021-9991. DOI: 10.1016/j.jcp.2019.06.032 (cit. on pp. xxii, 7).

Wu, Xie, et al. (Apr. 26, 2021). "A Comprehensive Survey of Inverse Uncertainty Quantification of Physical Model Parameters in Nuclear System Thermal-Hydraulics Codes". arXiv: 2104.12919 [stat] (cit. on pp. xxii, 7).

Xu et al. (Apr. 28, 2014). "A Robust Error-Pursuing Sequential Sampling Approach for Global Metamodeling Based on Voronoi Diagram and Cross Validation". In: *Journal of Mechanical Design* 136.7. ISSN: 1050-0472. DOI: 10.1115/1.4027161 (cit. on p. 26).

Ying (Feb. 1, 1991). "Asymptotic Properties of a Maximum Likelihood Estimator with Data from a Gaussian Process". In: *Journal of Multivariate Analysis* 36.2, pp. 280–296. ISSN: 0047-259X. DOI: 10.1016/0047-259X(91)90062-7 (cit. on p. 62).

Yoo, Estrada-Perez, and Hassan (Sept. 1, 2016). "Experimental Study on Bubble Dynamics and Wall Heat Transfer Arising from a Single Nucleation Site at Subcooled Flow Boiling Conditions – Part 2: Data Analysis on Sliding Bubble Characteristics and Associated Wall Heat Transfer". In: *International Journal of Multiphase Flow* 84, pp. 292–314. ISSN: 0301-9322. DOI: 10.1016/j.ijmultiphaseflow.2016.04.019 (cit. on p. 102).

Yun et al. (Dec. 1, 2010). "Experimental Investigation of Local Two-Phase Flow Parameters of a Subcooled Boiling Flow in an Annulus". In: *Nuclear Engineering and Design*. The 6th Japan-Korea Symposium on Nuclear Thermal Hydraulics and Safety - NTHAS6 Special Section 240.12, pp. 3956–3966. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2010.02.004 (cit. on p. 129).

Zhang and Wang (2010). "Kriging and Cross-Validation for Massive Spatial Data." In: *Environmetrics* 21.3/4, pp. 290–304. ISSN: 1180-4009 (cit. on p. 28).

Zhang (Mar. 2004). "Inconsistent Estimation and Asymptotically Equal Interpolations in Model-Based Geostatistics". In: *Journal of the American Statistical Association* 99.465, pp. 250–261. ISSN: 0162-1459, 1537-274X. DOI: 10.1198/016214504000000241 (cit. on p. 62).

Zhao and Mousseau (Aug. 1, 2012). "Use of Forward Sensitivity Analysis Method to Improve Code Scaling, Applicability, and Uncertainty (CSAU) Methodology". In: *Nuclear Engineering and Design*. The 8th International Topical Meeting on Nuclear Thermal-Hydraulics, Operation and Safety (NUTHOS-8 249, pp. 188–196. ISSN: 0029-5493. DOI: 10.1016/j.nucengdes.2011.09.042 (cit. on pp. xxii, 7).

ECOLE
DOCTORALE
DE MATHEMATIQUES
HADAMARD

Titre : Inférence bayésienne de l'erreur de modèle pour la calibration des codes de mécanique des fluides numérique diphasique

Mots clés : Quantification d'Incertitudes, Calibration Bayésienne, Erreur de Modèle, Processus Gaussiens, Mécanique des Fluides Numérique.

Résumé : Cette thèse concerne la calibration de codes de calcul, avec des applications dans le domaine nucléaire. Nous étudions des situations où il existe une distance irréductible entre les observations d'un phénomène physique et les prédictions obtenues par le modèle numérique, ce qui indique la présence d'une erreur de modèle. Cette erreur peut être estimée dans un cadre bayésien, pour lequel les paramètres du code sont représentés par des variables aléatoires. Nous proposons une nouvelle approche d'estimation de l'erreur de modèle, conduisant à l'introduction d'une nouvelle paramétrisation de la distribution du terme de biais de modèle. Nous démontrons sous l'hypothèse de normalité de la densité à posteriori la supériorité de notre approche vis-à-vis de la méthode classique car elle ne sous-estime pas l'incertitude paramétrique et révèle la totalité des différentes explications possibles des données. Nous étudions ensuite la validité de notre hypothèse de construction dans des cadres asymptotiques à grand nombre d'observations. Nous proposons un algorithme itératif pour la construction efficace d'un méta-modèle de la distribution du terme de biais. Les temps de calcul nous permettent de considérer des modèles d'erreur avec un grand nombre d'hyperparamètres, que nous appliquons à la calibration d'un modèle de répartition de flux de chaleur en ébullition. Nous démontrons ensuite la robustesse de notre approche sur la calibration du modèle de transport de l'aire interfaciale dans le code Neptune_CFD, en utilisant des données des expériences DEBORA. La robustesse de la méthode la rend applicable à de nombreux domaines au-delà du nucléaire.

Title : Bayesian inference of model error for the calibration of CFD codes

Keywords : Uncertainty Quantification, Bayesian Calibration, Model Error, Gaussian Processes, Computational Fluid Mechanics.

Abstract : In this thesis, we tackle the calibration of computer codes with applications in nuclear engineering. The focus is on situations where there is an irreducible discrepancy between model predictions and measurements, which reveals the presence of model error. This error can be estimated in a Bayesian setting, treating model parameters as random variables. Our approach introduces an innovative estimator of model error, which leads to a new parameterization of the model discrepancy distribution. Under the assumption of normality of the posterior, this provides an accurate determination of the parameter uncertainty and reveals all possible explanations of the data, thus proving superior to the traditional calibration framework. We examine the validity of the normality assumption under three asymptotic regimes with an increasing number of observations. We propose an adaptive design of computer experiments to build an accurate surrogate model of the model discrepancy distribution to reduce the numerical cost of calibration. This reduction allows us to calibrate a wall heat flux partitioning model with a distributed discrepancy model over multiple experimental configurations. The robustness of our approach is demonstrated with the calibration of the interfacial area transport model in the NEPTUNE_CFD code, using data from the DEBORA experiment. This framework applies to many problems, and its robust character makes it adequate for nuclear applications.

Institut Polytechnique de Paris
91120 Palaiseau, France