Instantaneous turbulent kinetic energy modelling based on Lagrangian stochastic approach in CFD and application to wind energy

Mireille Bossy, Jean-Francois Jabir, Kerlyns Martinez Rodriguez

Version of Record: https://www.sciencedirect.com/science/article/pii/S002199912100824X
Manuscript_3f8f55c799c653966e5d4c7f25d753e1

Instantaneous turbulent kinetic energy modelling based on Lagrangian stochastic approach in CFD and application to wind energy

Mireille Bossy[a], Jean-François Jabir[b], Kerlyns Martínez Rodríguez[a,1]

[a] *Université Côte d'Azur, Inria, CNRS, Sophia-Antipolis, France*
[b] *HSE University, Moscow, Russia Federation*
[c] *Institute of Statistics, University of Valparaíso, Valparaíso, Chile*

Abstract

We present the construction of an original stochastic model for the instantaneous turbulent kinetic energy at a given point of a flow, and we validate estimator methods on this model with observational data examples. Motivated by the need for wind energy industry of acquiring relevant statistical information of air motion at a local place, we adopt the Lagrangian description of fluid flows to derive, from the 3D+time equations of the physics, a 0D+time-stochastic model for the time series of the instantaneous turbulent kinetic energy at a given position. Specifically, based on the Lagrangian stochastic description of a generic fluid-particles, we derive a family of mean-field dynamics featuring the square norm of the turbulent velocity. By approximating at equilibrium the characteristic nonlinear terms of the dynamics, we recover the so called Cox-Ingersoll-Ross process, which was previously suggested in the literature for modelling wind speed. We then propose a calibration procedure for the parameters employing both direct methods and Bayesian inference. In particular, we show the consistency of the estimators and validate the model through the quantification of uncertainty, with respect to the range of values given in the literature for some physical constants of turbulence modelling.

Keywords: Wind energy dynamical model, Stochastic differential equation, Calibration, Lagrangian models, Turbulent kinetic energy, Uncertainty quantification.

Contents

Email addresses: mireille.bossy@inria.fr (Mireille Bossy), jjabir@hse.ru (Jean-François Jabir), kerlyns.martinez-rodriguez@inria.fr (Kerlyns Martínez Rodríguez)

1. Introduction

The need of statistical information on the wind, at a given location and on large time period, is of major importance in many applications such as the structural safety of large construction projects or the economy of a wind farm, whether it concerns an investment project, a wind farm operation or its repowering. The evaluation of the local wind is expressed on different time scales: monthly, annually or over several decades for resource assessment, daily, hourly or even less for dynamical forecasting (these scales being addressed with an increasing panel of methodologies, see e.g. [19]). In the literature, wind forecasting models are generally classified into physical models (numerical weather prediction models), statistical approaches (time-series models, machine learning models, and more recently deep learning methods), and hybrid physical and statistical models, see e.g. [52, 19, 31]. At a given site and height in the atmospheric boundary layer, measuring instruments (anemometer mast or nacelle anemometer) record time series of characteristics of the wind. For instance, wind speed characterizing load conditions, wind direction, kinetic energy and possibly power production. Such observation should feed into forecasting, but also uncertainty modelling. These observations present various time scales, some large ones such as diurnal, weekly and seasonal changes and variations, and some small scales often referred to as atmospheric turbulence [36]. In this context, probabilistic or statistical approaches are widely used, helping to characterize uncertainty through quantile indicators.

Recently, dynamical diffusion models have been proposed in the literature, featuring a continuous time description of wind dynamics. In [6], the Cox, Ingersoll and Ross (CIR) stochastic process –originally introduced in mathematical finance to model short term interest rates – is proposed to describe the dynamics of the squared wind speed. In [1, 2], an Ornstein-Uhlenbeck (OU) process is proposed and combined with a statistical measure of atmospheric turbulence called turbulent intensity.

In this paper, we derive a statistical model for the local wind speed, obtaining a reduced 0D+time equation from the 3D+time averaged Navier-Stokes equation with subgrid turbulence model. More precisely, we derive a continuous-time stochastic diffusion model for the instantaneous wind speed fluctuation called instantaneous turbulent kinetic energy measured at a given point. For that purpose, we start from the physical description of the fluid flow in a Lagrangian formulation that represents the fluid-particle dynamics with a stochastic diffusion process.

Turbulent flow models. Airflow is described by the incompressible Navier-Stokes equations. In the atmospheric boundary layer (ABL), the vicinity of the earth's surface, many vortex scales are present and interact. The flow is then described as turbulent, a situation that complexes any prediction by numerical approach. In particular, this makes the direct numerical approach (DNS) impossible, requiring to refine the spatial discretisation scales below the Kolmogorov scale (the scale from which viscosity allows to dissipate the kinetic energy, of a few millimetres in the atmospheric boundary layer). Common computational models are based on statistical approximations that replace unsolved subgrid scales, and on the idea that numerical estimations on averaged quantities of the flow (rather than in the details of the fluctuations) concentrated the major interest in many applications. Among the most well-known averaging methods, the Reynolds averaging of the Navier-Stokes equations –leading to the RANS models– use a statistical averaging decomposing each instantaneous variable of a flow into the sum of a mean part and a fluctuating part:

$$u' = U - \langle U \rangle$$

where U is the velocity field, $\langle U \rangle$ is the mean velocity (or average velocity) and u' represents the fluctuating or turbulent velocity. Averaging the Navier-Stokes equations has the effect of removing the fast fluctuation terms, but introduce unclosed term such as turbulent velocity covariance known as the Reynolds stress tensor $\langle\langle u'^{(i)} u'^{(j)} \rangle\rangle$. The question of *what might be a good model for this tensor ?* has been subject to a great interest and abundant literature (see e.g. in [48]), in the context of computational fluid dynamics (CFD). Before introducing classical closure models used in the ABL, we first present the Lagrangian point of view of the averaged Navier-Stokes equation, corresponding to the starting point of the stochastic reduced model analysed in this paper.

Fluid-particle based models. Stochastic Lagrangian approaches for turbulent flow were introduced at first for laden turbulent flow, in order to represent the turbulent subgrid-scale fluctuations of particle velocities that cannot be adequately resolved by mesh computation alone [47]. These approaches are used in the case of disperse two-phase flows where one phase is a set of discrete elements or 'particles' [45, 40] and they have been implemented in various complex industrial applications [41]. In the context of atmospheric flow, the so-called Lagrangian Particle Dispersion Models (LPDMs) are widely used for the analysis of air pollutants dispersion (see [53] and the references therein). Such methods adopts the perspective of an 'air parcel' by tracking a number of fictitious (or statistical) particles with position X_t released into a flow field:

$$dX_t = \langle U \rangle(t, X_t)dt + u'(t)dt, \tag{1.1}$$

where $u'(t)$ is a random fluctuation of the particle velocity, $\langle U \rangle$ is the mean velocity computed on a mesh. The velocity fluctuation is modelled using a stochastic differential equation (SDE) the complexity of which varies with the number of physical variables to be represented. But restricted to the turbulent velocity, it is generally declined with the simple Langevin model:

$$du'(t) = -\frac{u'(t)}{T_L}dt + \sqrt{C_0\varepsilon(t, X_t)}dB_t, \tag{1.2}$$

where the stochastic (or fast) part of the motion is described by the 3-dimensional Brownian motion B, amplified with the turbulent pseudo dissipation of the flow ε.

From a modelling view point, a great interest of the Lagrangian approach is its ability to represent the mean velocity and the Reynolds stress tensor as the first and second moments of the probability density function (PDF) of the Lagrangian model. A PDF method for turbulent flow[1] (firstly introduced by Pope [45]) considers stochastic processes $((X_t, U_t); 0 \leq t \leq T)$, describing the instantaneous position $X_t \in \mathbb{R}^3$ of a fluid particle and its instantaneous velocity $U_t \in \mathbb{R}^3$. Therefore, denoting the joint probability density of the process $((X_t, U_t); 0 \leq t \leq T)$ by ϱ, for all suitable map $g : v \in \mathbb{R}^3 \mapsto g(v) \in \mathbb{R}$, the Reynolds operator $\langle \cdot \rangle$ applied to $g(U)$ is interpreted as the probabilistic conditional expectation of the particle velocity U_t, knowing that its position X_t is x, under the probability $\mathbb{P}$ of the model provided with expectation symbol $\mathbb{E}$:

$$\langle g(U) \rangle(t, x) = \mathbb{E}\left[g(U_t) |\ X_t = x\right] = \frac{\int_{\mathbb{R}^3} g(u)\varrho(t, x, u)du}{\int_{\mathbb{R}^3} \varrho(t, x, u)du}. \tag{1.3}$$

Macroscopic quantities of interest can be then identified with this rule, as long as the chosen Reynolds stress closure order allows to represent them in the model (we refer to Section 2 for more details).

For the use of the fluid-particle method in meteorological context, we particularly refer the interested reader to [4, 54] for application to filtering of wind data, [9, 10] for application to refined wind computation, and more recently [14, 15] for wind farm simulation.

Aim and layout of the paper. In this paper, starting from the 3D+time physical description of the flow given by the stochastic Lagrangian model framework, we construct a simplified 0D+time model featuring the evolution of the turbulent kinetic energy of the flow at a fixed location. In addition we construct and analyse a calibration procedure that includes uncertainty as a part of the calibration methodology. We test the model and its calibration on wind measurements obtained from a typical measure mast used in wind energy potential assessment (a 30 m height mast, with a frequency of 10 Hz) on a large interval of time (a year) to capture a larger part of the variability of the wind during a seasonal cycle. We validate the method, comparing the posterior distribution of the model-parameter issued from the G_{ij} tensor coefficient in (2.2) below against the interval of the values proposed in the CFD literature. Next, we evaluate the ability of the calibrated model to replicate the observed time series. Finally, we come back on the calibration procedure, considering the 10-minutes-averaged turbulence intensity as a good statistical prediction value for the turbulent source term in the proposed stochastic model (3.9). In particular this method allows to well approximate the Weibull distribution form of the turbulent wind speed $\|u'\|$, showing the consistency of this simpler variant of the calibration method and the reliability of the model to predict efficiently the wind distribution.

In Section 2, we set the theoretical basis of the reduced model by shortly introducing the fluid-particle Lagrangian model in turbulent (near wall) flow. In Section 3, we mathematically derive in several steps the 0D+time model for the instantaneous turbulent kinetic energy (TKE). We analyse the wellposedness of the SDE describing the instantaneous TKE process (existence and trajectorial uniqueness). We give semi-explicit form of the moments

[1] We consider here only the case of constant mass density flow, for the sake of clarity.

solution and analyse their long-time behaviour, in order to further simplify the reduced model. We also present the wind speed observations targeted by our instantaneous TKE model (see Subsection 3.4) and used to calibrate and to validate the model in the next sections. We formulate the calibration procedure for the instantaneous TKE process in Section 4 in two main steps. Based on the mathematical analysis of the model, we first derive a pseudo-likelihood maximisation procedure (Step zero in 4.1), we next improve the calibration process, introducing a Bayesian procedure to handle the assumed uncertainty on the parameters (Step one in 4.2). Finally, in Section 5 we summarize the key findings and results of the calibration along with the validation of the model.

We end this introduction by motivating the calibration under uncertainty methodology that we promote in this paper.

Calibration with uncertainty for turbulent flow models. As pointed out in [44], to produce accurate predictions of quantities of interest it is necessary to make a systematic treatment of the uncertainties within the models and observations, quantifying them along their propagation through a computational model. In particular, estimation of model parameters comes before assessing model performance.

For instance in [27], studying the uncertainty of the RANS model parameters based on a Launder-Sharma turbulence closure relation, the authors used a Bayesian calibration method employing measured boundary-layer velocity profiles. By modelling the spread of parameters within the flow-class, they show the ability of the Bayesian calibration to provide information about the values these parameters should take in each flow case. This uncertainty is thus highlighted by its quantified distribution, suggesting that the parameters must be seen as tuned to be associated with the k-ε turbulence closure, and in general, parameters are not expected to be flow-independent.

Given the stochastic nature of Lagrangian modelling handled in this paper, the uncertainty quantification for the model parameters can be less costly to implement through the computation/approximation of densities. In particular the need for a surrogate model can be more easily mitigated. More precisely, considering a (deterministic) model, a set of observations, and the set of model parameters to estimate from this data, the obvious idea is to minimize one of the performance measures with respect to the parameter values. Instead, with a stochastic model, a more statistically reliable method of parameter estimation can be used, since (in principle) it is possible to construct the likelihood function which is merely the probability that the model generates the observed data, given a parameter set. Once a model for parameters uncertainty is identified, by modifying/increasing the variables dependency of the likelihood function, we may quantify the probability distribution of the parameters given the observational data, by the well-known method of maximum likelihood estimation. Note that the explicit computation of the likelihood requires to work with an explicit form for the density function, which drastically restricts the class of possible stochastic models. Alternative approximation methods are available, based on discrete time sampling such as Markov Chain Monte Carlo (MCMC) methods that we detail in Section 4 and implement in Section 5.

2. A short review of stochastic Lagrangian approach for turbulent flows

In this section we introduce the framework for the stochastic Lagrangian approach (also known as PDF approach) describing the dynamics of fluid particles within turbulent flows. Ideally, a fluid particle is a tracer that moves with the local flow velocity. Considering $X^+(t,Y)$ the position at time t of a tracer initially located at the point Y -called Lagrangian coordinate, or material coordinate- at a specified time t_0, the tracer evolves according to $\frac{\partial}{\partial t}X^+(t,Y) = U(t,X^+(t,Y))$ where $(t,x) \mapsto U(t,x)$ is here the Eulerian velocity field. The Lagrangian velocity U^+ is then defined in terms of its Eulerian counterpart

$$U^+(t,Y) = U(t,X^+(t,Y)).$$

In CFD, turbulence modelling gives access only to the averaged Eulerian velocity and other second moments according to the model. Stochastic Lagrangian approach focus on describing the dynamics of a fluid-particle -or virtual fluid parcel- and its characteristic position and instantaneous velocity (X_t,U_t), dynamics characterized by a SDE which suitably approximate the motion of $(X^+(t),U(t,X^+(t)))$. This SDE is constructed on the basis of a transport equation (or Fokker Planck equation) for the density function relative to the position and the velocity of the fluid particle. This joint probability density of the process $((X_t,U_t);0\leq t\leq T)$, denoted below by ϱ, allows to interpret the Reynolds operator $\langle\cdot\rangle$ as in (1.3), the expectation symbol $\mathbb{E}$ being notably associated to the probability measure $\mathbb{P}$, under which the Brownian motion (B_t) driving the SDE is defined.

A reference stochastic Lagrangian model is the *Generalized Langevin Model* (GLM),

$$\begin{cases} dX_t^{(i)} = U_t^{(i)} dt,\ 1 \le i \le 3, \\ dU_t^{(i)} = -\frac{1}{\rho}\frac{\partial \langle P\rangle}{\partial x_i}(t, X_t)dt + \sum_{j=1}^{3} G_{ij}(t, X_t)(U_t^{(j)} - \langle U^{(j)}\rangle(t, X_t))dt + \sqrt{C_0(t, X_t)\varepsilon(t, X_t)}dB_t^{(i)}, \\ \langle P\rangle \text{ is the mean pressure, } (B_t = (B_t^{(i)}),\ t \ge 0) \text{ is a standard 3D-Brownian motion,} \end{cases} \tag{2.1}$$

designed to be consistent with the Navier-Stokes equations through formal developments on the Fokker-Planck equation derived from (2.1). This reference model involves a generalized return-to-equilibrium tensor $G_{ij}(U^{(j)} - \langle U^{(j)}\rangle)$ and the dissipation rate ε defined by

$$\varepsilon(t, x) = \tfrac{1}{2}\nu \sum_{i,j=1}^{3} \langle (\partial_{x_j} u'^{(i)} + \partial_{x_i} u'^{(j)})^2\rangle(t, x),$$

with ν the kinematic viscosity of the fluid.

The probability measure supporting the Brownian motion B and the solution of the GLM equation (2.1) allow us to define the mean velocity field of the flow as the conditional expectation

$$\langle U^{(i)}\rangle(t, x) = \mathbb{E}[U_t^{(i)} | X_t = x],$$

and to define the Lagrangian turbulent part of the instantaneous velocity:

$$u'_t = U_t - \langle U\rangle(t, X_t).$$

Then the stochastic Lagrangian equation for the process $((X_t, u'_t);\ t \ge 0)$ is

$$\begin{cases} dX_t^{(i)} = (u_t'^{(i)} + \langle U^{(i)}\rangle(t, X_t))dt,\ 1 \le i \le 3, \\ du_t'^{(i)} = \displaystyle\sum_{j=1}^{3} G_{ij}(t, X_t)u_t'^{(j)}dt + \sqrt{C_0(t, X_t)\varepsilon(t, X_t)}dB_t^{(i)}, \end{cases} \tag{2.2}$$

where the dimensionless coefficient $C_0(t, x)$ may be a constant or may depend on the local values of the Reynolds stress tensor, the dissipation rate and the drag force $\partial_{x_j}\langle U^{(i)}\rangle$.

For simulation purposes (see for instance [10] for a general presentation and application to downscaling methods, [14, 15] for wake turbine and wind farm simulation), a discrete time particle method can be applied to stochastic Lagrangian models to obtain an approximation for the mean flow components. The main difficulty of these approximation methods regards the non-linearity driven by the conditional expectation, and can be overcome using a kernel-regularization approach and particles in cell methods [10, 15].

In the literature, several models with distinct referential values for the parameters exist. Among them, we mention the Rotta model related to simplified Langevin model (SLM), the isotropization-to-production model (IPM), also the Shih-Lumley model (SL), and the Launder-Reece-Rodi Model (LRR). We further refer the interested reader to [46] to a detailled presentation of these -and other- models.

2.1. Models for the tensor G

We present a selection of possible models, chosen for their use in the ABL modelling, characterized by high-Reynolds number and anisotropic flows.

The Simplified Langevin model (SLM) . Proposed by Pope [45], and later seen as a particular case of GLM [48], the SLM assumes the tensor G as an isotropic (diagonal) tensor given by

$$G_{ij}(t, x) = -\left(\frac{1}{2} + \frac{3}{4}C_0\right)\frac{\varepsilon}{\mathrm{k}}(t, x)\ \delta_{ij}, \tag{2.3}$$

where δ_{ij} is the Kronecker delta[2] and $\mathrm{k}(t, x)$ is the mean turbulent kinetic energy defined by

$$\mathrm{k}(t, x) = \tfrac{1}{2}\sum_{i=1}^{3} \mathbb{E}[(u_t'^{(i)})^2 | X_t = x]. \tag{2.4}$$

[2] $\delta_{ij} = 1$ if $i = j$, 0 otherwise.

It has been shown (e.g. [48]) that the model tensor (2.3) is consistent with the simple Rotta's return-to-isotropy model for the dynamics of the Reynolds stresses $\langle u'^{(i)}u'^{(j)}\rangle$ (so it is called simple). The condition for this consistency is to identify[3] the so-called Rotta's constant C_R with

$$C_R = 1 + \tfrac{3}{2}C_0. \tag{2.5}$$

Notice that, the constant C_0 in the model (2.2) must be non-negative in order to have a well-defined diffusion term. Thus, in view of (2.5), C_R must satisfied the Rotta condition, $C_R \geq 1$ [26].

The isotropization-to-production (IP) model for homogeneous turbulence. This GLM version is constructed to be consistent with the LRR-IP model [47, 48] (see also [26, 22]):

$$G_{ij}(t,x) = -\frac{C_R}{2}\frac{\varepsilon}{\mathrm{k}}(t,x)\delta_{ij} + C_2\partial_{x_j}\langle U^{(i)}\rangle(t,x). \tag{2.6}$$

The consistency with the LRR-IP model is ensured by the choice of a diffusion coefficient computed from the turbulent production tensor $\mathcal{P}_{ij}$ (and coming with a realizability constraint [26]) :

$$C_0\,\varepsilon(t,x) = \tfrac{2}{3}\left(C_R\,\varepsilon(t,x) + C_2\,\mathcal{P}(t,x) - \varepsilon(t,x)\right), \tag{2.7}$$

where, adopting Einstein notation, $\mathcal{P}(t,x) = \frac{1}{2}\mathcal{P}_{ii}(t,x)$ is the turbulent production derived from

$$\mathcal{P}_{ij} := -\sum_{l=1}^{3}\langle u'^{(i)}u'^{(l)}\rangle\partial_{x_l}\langle U^{(i)}\rangle + \langle u'^{(j)}u'^{(l)}\rangle\partial_{x_l}\langle U^{(j)}\rangle.$$

Notice that C_0 in the definition of the tensor G for the IP model (2.7) is no more considered as constant. Notice also that the realizability constraint requiring C_0 to be positive reads as $C_R \geq 1 - C_2\frac{\mathcal{P}}{\varepsilon}$. The SLM is a particular case of the IP model considering $C_2 = 0$.

Elliptic blending model. The purpose of this model is to add an anisotropic effect near the ground (or wall effect, see e.g. [24, 25, 38, 56]). It starts with the following form of G and diffusion coefficient:

$$G_{ij} = -\gamma_{ij} - \frac{1}{2}\frac{\varepsilon}{\mathrm{k}}\delta_{ij}, \qquad C_0\,\varepsilon = \sum_{i,j=1}^{3}\frac{2}{3}\gamma_{ij}\langle u'^{(i)}u'^{(j)}\rangle, \tag{2.8}$$

where the tensor γ is 'interpolated' between a near wall Reynolds stress model and the homogeneous model already described [25]. In its simplified version [38],

$$\gamma_{ij}(t,x) = (1-\alpha(t,x)\mathrm{k}(t,x))\gamma^w_{ij}(t,x) + \alpha(t,x)\mathrm{k}(t,x)\gamma^h_{ij}(t,x),$$

where $\alpha(t,x)$ is the elliptic blending coefficient given as the solution of the elliptic partial differential equation:

$$L^2\nabla^2\alpha(t,x) - \alpha(t,x) = -\frac{1}{\mathrm{k}}(t,x),$$

with L being a characteristic length scale (possibly varying in time and space). We refer to [56] and the references therein for the modelling of γ^w_{ij} and γ^h_{ij}. Notice that, here again, C_0 is no more considered as constant and depends in particular on α.

Remark 2.1. *It should be noted that in both Lagrangian and Eulerian approaches, the values of the coefficients C_0 and C_R, known as the Kolmogorov constant and the Rotta constant, might vary according to the model and context. For instance, the values $C_0 = 2.1$ and $C_R = 4.15$ are suggested in [46]. Nevertheless, the IP model (see Equation (2.7)) may assume $C_R = 1.5$ or $C_R = 1.8$, [26]. Similarly for the LRR model, $C_R = 1.5$, [46]. From [46, Appendix], we quote that $C_R = 1.0$ corresponds to no-return-to-isotropy, while values from 1.5 to 5.0 have been suggested by different authors. This implies that the values of C_0 might vary between $\frac{1}{3}$ and $\frac{8}{3}$ (using the identification (2.5)).*

[3] A more general relationship (see [26, pp. 398]) is given by $C_0 = \frac{2}{3}(\frac{\mathrm{k}}{\tau\varepsilon}C_R - 1)$ for τ some characteristic time scale. In the case of Rotta's closure we have $\tau = \varepsilon/\mathrm{k}$.

2.2. *Parameterization of the dissipation rate*

The GLM or its Fokker Planck equation contains no information on the turbulence timescale –unless thought the addition of a direct dissipation state variable [48, 12.5]– and needs to be supplemented with a model for the dissipation rate of the kinetic energy. We present below only two simple models based on the relation $\mathrm{k}-\varepsilon$, but refinements are proposed in the literature.

The mixing length parameterization. This relation is classically used in the ABL:

$$\varepsilon(t,x) = \frac{C_\varepsilon}{\ell_{\mathrm{m}}}\mathrm{k}^{\frac{3}{2}}(t,x), \tag{2.9}$$

where C_ε is a constant to be chosen, ℓ_{m} is a characteristic length scale called *m*ixing length. In general, near the ground, the value of the mixing length ℓ_{m} does vary linearly with respect to κz, where z denotes the distance to the wall from x, and κ is the Von Kármán constant (see [9] and the references therein for further discussion).

The turbulent viscosity parameterization. In the $\mathrm{k}-\varepsilon$ model framework [48], the turbulence eddy viscosity is given by $\nu_{\text{turb}} = C_\mu \frac{\mathrm{k}^2}{\varepsilon}$ with $C_\mu = 0.09$. In addition, near the wall (behind a height of 150 m, to give an order of magnitude in the ABL) ν_{turb} is related with the velocity friction thought $\nu_{\text{turb}}(z) = u^*\kappa z$, leading to the following parameterization

$$\varepsilon(t,x) = \frac{C_\mu}{u^*\kappa z}\mathrm{k}^2(t,x). \tag{2.10}$$

The velocity friction u^* depends on the flow and on the ground (the roughness length of the terrain among others factors).

3. Reduced Lagrangian model for the instantaneous turbulent kinetic energy

As already mentioned in the introduction, modelling characteristics of the wind at a fixed location is of great interest for many applications. In this context, fluid-particle based turbulent flow models offer a simple way to reduce from '3D+time' dependent model to 'time' model. For a Lagrangian stochastic model, the '3D' field notion is a conditional expectation that can be approximated at a fixed location by a calibration procedure. As the step in this direction, we refer to Baehr [4] which was interested in a filtering methodology for one point (eventually mobile) of wind observation and which proposed to use local random model such as Lagrangian models combined with nonlinear filtering techniques to clean wind measurements. These ideas were applied later to Doppler wind LIDAR observations [54, 5, 50].

In our case, we assume (filtered) data available at a fixed location x_{obs} (see Section 3.4), ready to be used to calibrate a time-model. The idea developed in this paper is to somehow reduce the stochastic Lagrangian modelling into instantaneous turbulent kinetic energy as a time-model, while considering the uncertainty of the physical parameters of the model. Renewable energy development has raised growing interest in energy production forecasting and modelling. Several approaches are taking into account the uncertain nature of the forecast through stochastic modelling. For recent examples based on stochastic diffusion models, we refer for instance to [3, 42] for the construction of probabilistic forecast of solar irradiance, with simple linear drift form. In [3], a diffusion process is proposed, resulting from a deterministic forecast, and the parameters involved are estimated by means of a variance-autocorrelation fitting. The diffusion coefficient used to model the solar irradiance has the power form $x^\alpha(1-x)^\alpha$, with the constant α to be calibrate from data. Concerning wind energy production, [2] has proposed a data-driven Ornstein-Uhlenbeck model describing the wind speed on a scale of seconds. On the other hand, the Weibull distribution has been widely used in wind energy and other renewable energy sources [17], where the main issue has been the estimation of the distribution coefficients. Based on this, in [6], a stochastic model of the squared norm of the wind velocity as a CIR process have been proposed with coefficients to be calibrate. These models have in common the fact that they are inspired from some a priori on their parametric forms.

In this section we derive a physical-based model describing the instantaneous TKE localized at a fixed location x_{obs}.

3.1. *First step: localized dynamics of the norm of the turbulent velocity*

We start from the SLM equations (2.2,2.3), and applying the Itô lemma, we obtain a first equation for the squared-norm of the turbulent velocity that we force to be localized at x_{obs}

$$d\|u'_t\|^2 = -2\left(\frac{1}{2}+\frac{3}{4}C_0\right)\frac{\varepsilon}{\mathrm{k}}(t,x_{\text{obs}})\|u'_t\|^2dt + 3C_0\varepsilon(t,x_{\text{obs}})dt + 2\sqrt{C_0\varepsilon(t,x_{\text{obs}})}\sum_i u'^{(i)}_t dB^{(i)}_t. \tag{3.1}$$

Equation (3.1) is a version of the fluid-particle model (2.2,2.3), conditioned at each time t by the event that the fluid-particle position X_t is going through the point x_{obs}. We proceed by defining the *instantaneous turbulent kinetic energy* at x_{obs} as the stochastic process $(q_t;\ t \geq 0)$ given as the formal solution to (3.1). In particular, we get the relation

$$\mathrm{k}(t, x_{\text{obs}}) = \frac{1}{2}\mathbb{E}[q_t].$$

In order to simplify notation, hereafter, k_t (respectively ε_t) will denote the mean turbulent kinetic energy at the fixed position $\mathrm{k}(t, x_{\text{obs}})$ (respectively $\varepsilon(t, x_{\text{obs}})$). According to the Levy's characterization Theorem of Brownian motion (see e.g. [49]), the process $W_t = \sum_i \int_0^t \frac{u_s'^{(i)}}{\|u_s'\|} dB_s^{(i)}$ identifies as a one-dimensional Brownian motion. Further, assuming that the initial turbulent energy q_0 is known, we obtain the following SDE for the dynamics of the process $(q_t; t \geq 0)$:

$$dq_t = -C_R \frac{\varepsilon_t}{\mathrm{k}_t} q_t dt + 3C_0 \varepsilon_t dt + 2\sqrt{C_0 \varepsilon_t}\sqrt{q_t} dW_t, \quad q_0 \text{ given.} \tag{3.2}$$

For this first SDE obtained from SLM, all that remains to be specified a priori is the dissipation ε_t. However, in wind energy application context, the fluid particle model is used in the vicinity of the ground, and anisotropic effects cannot be simply neglected, as well as buoyancy effects due to temperature variation. In the one hand, these effects are complex to model. On the other hand, the mean long time estimate obtained in Lemma 3.4 below shows that the behaviour of the SLM alone is the systematic return to zero turbulence, in contradiction with the observations (see Section 3.4). This behaviour evidences the need of an additional source term in the SDE accounting for the regime of turbulence production near the wall. In order to replicate this effect, we consider an extra drift term γ accounting for the extra-diagonal (non-isotropic) contributions of the tensor G_{ij} not retained in the SLM:

$$dq_t = \gamma\ dt - C_R \frac{\varepsilon_t}{\mathrm{k}_t} q_t\ dt + 3C_0 \varepsilon_t dt + 2\sqrt{C_0 \varepsilon_t}\sqrt{q_t} dW_t, \quad q_0 \text{ given.} \tag{3.3}$$

In Section 4, we first propose a calibration procedure for γ, updated in time according to an intermediate time scale, and next a higher one. We also propose to use the (renormalized) wind turbulence intensity as a candidate to approximate γ. In order to simplify the discussion on the wellposedness analysis and the computation of the a priori estimators, in this first part we consider γ as constant (in the sense that the variation in time of γ is at a lower frequency than those of q_t).

3.2. Second step: incorporating the dissipation parametization

In order to close the term ε_t in the SDE (3.3), we choose the mixing length closure hypothesis (2.9) used in ABL modelling. For simplicity, we consider ℓ_{m} constant with $\ell_{\text{m}} = \kappa z_{\ell_{\text{m}}}$, where κ is the Von Kármán constant and $z_{\ell_{\text{m}}}$ is the height at which the measurements were taken:

$$\varepsilon_t = C_\alpha \mathrm{k}_t^{3/2} = \frac{C_\alpha}{2\sqrt{2}} \mathbb{E}^{3/2}[q_t],$$

where $C_\alpha := \frac{C_\varepsilon}{\kappa z(x_{\text{obs}})}$, and where we use here and later the notation $\mathbb{E}^\beta[q_t] = (\mathbb{E}[q_t])^\beta$ for any $\beta > 0$. Introducing this relation in Equation (3.3), we obtain the following CIR-type **stochastic mean-field TKE model**:

$$dq_t = \gamma dt - C_R \frac{C_\alpha}{\sqrt{2}} q_t \mathbb{E}^{1/2}[q_t] dt + 3C_0 \frac{C_\alpha}{2\sqrt{2}} \mathbb{E}^{3/2}[q_t] dt + \sqrt{\sqrt{2} C_0 C_\alpha}\ \mathbb{E}^{3/4}[q_t] \sqrt{q_t} dW_t, \quad q_0 \text{ given.} \tag{3.4}$$

Remark 3.1. *Formally, taking expectation on both sides of* (3.4) *and using* (2.5) *leads to the ordinary differential equation for* $\mathrm{k}_t = \mathbb{E}[q_t]/2$

$$\frac{d\mathrm{k}_t}{dt} = \left(\frac{\gamma}{2} - \varepsilon_t\right) \mathbb{1}_{\{\mathrm{k}_t > 0\}}, \quad \mathrm{k}_0 = \frac{q_0}{2}$$

which corresponds to the classical equation form for k*, involving the turbulence production terms (here reduced to* γ*) deduced from RANS equations with a* k-ε *closure model [26]. As the true production process is replaced by the model constant* γ*, the indicator function in the right hand-side prevents* k *to become non-positive. Coming back to the autonomous ODE form,*

$$\frac{d\mathrm{k}_t}{dt} = \left(\frac{\gamma}{2} - C_\alpha \mathrm{k}_t^{3/2}\right) \mathbb{1}_{\{\mathrm{k}_t > 0\}}, \quad \mathrm{k}_0 = \frac{q_0}{2}, \tag{3.5}$$

under the conditions that q_0 *and* γ *are non-negative, Lemma 3.4 ensures that a unique non-negative solution to* (3.5) *exists for all* $t \geq 0$.

The SDE (3.4) features two main characteristics. On the first hand, the SDE can be classified as of CIR type, since the dependence in q of the diffusion coefficient is in a square root form, a non-Lipschitz dependence well studied in the literature (see Appendix A.1 and [33]). On the second hand, the SDE can be also classified as of mean-field type, since both drift and diffusion coefficients depend nonlinearly on the law's process through the expectation term $\mathbb{E}[q_t] = 2\mathrm{k}_t$. The combination of these features categorizes the model as a McKean-Vlasov SDE (we refer the interested reader to the surveys [11] and [34] for some review on theoretical aspects related to these models and their numerical approximation). At first sight, justifying the existence of a global-in-time solution to the equation (3.4) is not trivial. While positiveness criteria for solution of SDEs with polynomial coefficients are well-understood, the dependency of the coefficients in fractional powers of $\mathbb{E}[q_t]$ makes the application of such criteria not immediate. Nevertheless, an opportune aspect of the model lies in the dynamics of $\mathbb{E}[q_t]$, which as mentioned in Remark 3.1 is driven by the autonomous ODE (3.5). We deduce the existence and trajectorial uniqueness of the positive solution to the SDE (3.4) from the a priori behaviour of k stated in Lemma 3.4.

Proposition 3.2. *Consider the positive parameter set C_α, C_0, and let $\gamma \geq 0$. Then, there exists a unique strong positive solution $(q_t;\ t \geq 0)$ to the McKean-Vlasov SDE (3.4).*

The proof of Proposition 3.2 is postponed to Appendix A.1.

Remark 3.3. *A negative production term $\gamma < 0$ is not forbidden in the model, if it is designed to decrease to zero with* k. *For the sake of simplicity we assume $\gamma \geq 0$ as a constant updated on time sub-periods, and let the extension to possible function $q \mapsto \gamma(q) \in \mathbb{R}$ to future improvement.*

On the dissipation of the energy.

Lemma 3.4. *Assume that there exists a solution $(q_t, t \geq 0)$ to (3.4). Then, for any $\gamma \geq 0$ the pth-moment is bounded (uniformly in time), for all $p \geq 1$, i.e. there exist some constants $\overline{C}, \underline{C} > 0$ independent on time such that*

$$\underline{C}\, \mathbb{1}_{\gamma>0} \leq \sup_{t\geq 0} \mathbb{E}[q_t^p] \leq \overline{C}. \tag{3.6}$$

In particular for $p = 1$, we have

$$\min\left\{q_0, \left(\frac{\sqrt{2}\gamma \mathbb{1}_{\{\gamma>0\}}}{C_\alpha}\right)^{2/3}\right\} \leq \sup_{t\geq 0} \mathbb{E}[q_t] \leq \max\left\{q_0, \left(\frac{\sqrt{2}\gamma \mathbb{1}_{\{\gamma>0\}}}{C_\alpha}\right)^{2/3}\right\}, \tag{3.7}$$

with long-time behaviour

$$\lim_{t\to+\infty} \mathbb{E}[q_t] = \left(\frac{\sqrt{2}\gamma \mathbb{1}_{\{\gamma>0\}}}{C_\alpha}\right)^{2/3}. \tag{3.8}$$

This lemma emphasises the role of the (non-dissipative) source term γ in the model. In particular when $\gamma > 0$, the instantaneous norm of the velocity fluctuation q_t has the non trivial limit (3.8). The proof is detailed in Appendix A.2.

3.3. Third step: deriving a CIR-like model for the instantaneous TKE

In this last step, we move from the McKean nonlinear model (3.4) to a linear one. The reason for this is that calibration methods for McKean processes are still little developed and limited to particular cases. Instead, we make use of the long-time convergence of the moment $\mathbb{E}[q_t]$ to simplify the dynamics (3.4). More precisely, from the nature of our observations (see Section 3.4 below), we assume that the production term γ is always positive in order to get a non zero long-time limit. Then, taking one by one the coefficients of (3.4)

$$\begin{aligned}
&\lim_{t\to+\infty} C_R \frac{C_\alpha}{\sqrt{2}} (\mathbb{E}[q_t])^{\frac{1}{2}} = C_R \left(\frac{C_\alpha^2 \gamma}{2}\right)^{1/3} := \Theta(C_\alpha, \gamma) \\
&\lim_{t\to+\infty} \gamma + 3C_0 \frac{C_\alpha}{2\sqrt{2}} (\mathbb{E}[q_t])^{\frac{3}{2}} = \gamma + \frac{3}{2} C_0 \gamma := \mu(C_\alpha, \gamma)\Theta(C_\alpha, \gamma) \\
&\lim_{t\to+\infty} \sqrt{\sqrt{2} C_0 C_\alpha}\, (\mathbb{E}[q_t])^{\frac{3}{4}} = \sqrt{2C_0\gamma} := \sigma(\gamma).
\end{aligned}$$

By doing so, we erase a part of the complexity of the transition to equilibrium, obtaining the following **CIR model for the instantaneous TKE**:

$$dq_t = \Theta(C_\alpha, \gamma)\left(\mu(C_\alpha, \gamma) - q_t\right) dt + \sigma(\gamma)\sqrt{q_t} dW_t, \quad q_0 \text{ given, and } \gamma > 0, \tag{3.9}$$

where

$$\Theta(C_\alpha,\gamma)=C_R\big(\frac{C_\alpha^2\gamma}{2}\big)^{1/3},\qquad \mu(C_\alpha,\gamma)=\big(\sqrt{2}\frac{\gamma}{C_\alpha}\big)^{2/3},\qquad \sigma(\gamma)=\sqrt{2C_0\gamma}.$$

Remarkably, starting from a well established turbulence model, and following the path of the reduction to time-model, we recover the class of diffusion models suggested in [6]. Our major contribution in this context is thus to provide an intrinsic physical meaning on the model parameters. Further, a more descriptive dynamics for the turbulent kinetic energy can be obtained from the non-linear model (3.4), where additional efforts are needed to propose an adequate calibration procedure. This last point will be the subject of future works.

A strong advantage of the model (3.9) is precisely that its solution, the CIR process [20], have been widely studied, although mainly in mathematical finance. For CIR processes, the parameter $\Theta > 0$ is the speed of adjustment to the mean μ, and the well-known assumption on the parameters

$$2\Theta(C_\alpha,\gamma)\mu(C_\alpha,\gamma)\geq\sigma^2(\gamma),$$

(from Feller's criteria [35, Theorem 5.5.29]) excludes the trajectories to visit zero, through a pushing upward effect of the drift against the diffusion when the process gets close to zero. In the particular case of Equation (3.9), this condition reads as

$$2C_R\geq 2C_0,$$

which, by the Rotta's relation (2.5), is always satisfied. Then, starting from a positive q_0, the uniqueness of a (strictly) positive trajectory satisfying Equation (3.9) is always guaranteed.

The CIR process is also an ergodic process with known stationary density. The CIR process has a well-known explicit relation with chi-squared random variables, which provides the transition density associated to the solution of Equation (3.9) in terms of Bessel functions. From this, the moments of the process $(q_t; t\geq 0)$ can be explicitly computed (see, e.g., [21]) with

$$\mathbb{E}[q_t^p]=\left(\frac{2\Theta}{\sigma^2\left(1-\exp\{-\Theta t\}\right)}\right)^{-p}\frac{\Gamma\left(2\mu\Theta/\sigma^2+p\right)}{\Gamma\left(2\mu\Theta/\sigma^2\right)}\,{}_1F_1\left(-p,\frac{2\mu\Theta}{\sigma^2};-\frac{2\Theta\,q_0}{\sigma^2(\exp\{\Theta t\}-1)}\right),\tag{3.10}$$

where Γ is the gamma function and ${}_1F_1$ is the confluent hypergeometric function[4] (here, we have been omitted the dependence on (C_α,γ) in the coefficients in order to simplify notations). Then, $\mathbb{E}[q_t^q]$ is finite for all $p>-\frac{2\mu\Theta}{\sigma^2}$.

Notice that the computation of the moments in (3.10) allows to recover the identity (3.8), underlying the consistency between the McKean nonlinear model (3.4) with its linear version (3.9).

3.4. Observational data

The data used in our study was obtained from the open observation platform of SIRTA[5] (Haeffelin et al. [30]). We used wind measurements taken at a mast of 30 meters height with a sonic anemometer. This instrument measures the wind components at a fixed single point, considered as an observer particle located at $z(x_{\text{obs}})$ =30 m. Precisely, the observational data is a family of time series containing, among other measures, the three components of the wind, registered with a frequency of 10 Hz during the year 2017. In order to take advantage of the great variability of wind situations represented in a so huge data set, without the inconvenience of processing and analysing such a large series for a summarized presentation, we have cut out the series, retaining only one day per week chosen arbitrarily and favouring daytime periods from 4 a.m. to 8 p.m (measured in local time). We thus have extracted a time series composed of 46 periods of 16 hours spanned for all Wednesdays of the year 2017. Our data set is then representative of the range of possible values for the temperature, degree of humidity, direction of the wind, intensity, and therefore a wide variety of wind profiles during the cycle of a year.

Given this time series $(U_t^{(i),\text{obs}}, i=1,2,3, t\geq 0)$, where t is incremented each $\frac{1}{10}$ seconds, we first extract from it the observed instantaneous TKE process $\boldsymbol{q}^{\text{obs}}=(q_t^{\text{obs}};\ t\geq 0)$ through the instantaneous turbulent velocity $(u_t';\ t\geq 0)$. In practice, it is very common in wind energy industry to approximate the mean velocity $\langle U\rangle$ by an average in time over an interval of 10 minutes to 60 minutes, corresponding to a minimum in the wind power spectral density. Hence, we compute:

$$q_t^{\text{obs}}\approx\Big\|U_t^{\text{obs}}-\frac{1}{\zeta}\sum_{t-\zeta\leq s<t}U_s^{\text{obs}}\Big\|^2,\qquad \text{for } t \text{ in the selected signal interval, } \Delta t=\tfrac{1}{10},\tag{3.11}$$

[4] also known as Kummer's function: ${}_1F_1(a,b,z)=\sum_{m=0}^{+\infty}\frac{a^{(m)}z^m}{b^{(m)}m!}$, with $a^{(m)}=a(a+1)\dots(a+m-1)$, the rising factorials.

[5] Site Instrumental de Recherche par Télédétection Atmosphérique.

with the time-window $\zeta = 40$ minutes. While seeking wind homogeneity periods, we may be interested in classify the observations in wind condition regimes. A way to do so is provided by the turbulence intensity (TI) measure of the wind, defined as (see e.g [32] and the reference therein) the quotient between the standard deviation of wind speed series and a representative mean velocity:

$$I_t := \frac{\sqrt{\langle q_t^{\text{obs}} \rangle}}{\sqrt{3} \| \langle U_{(d)}^{\text{obs}} \rangle \|}. \tag{3.12}$$

For low wind speeds, it is observed that high turbulence will enhance wind power performance. At the contrary when the wind speed is high and stable, high turbulence will lower the wind power production. In many references (e.g [32, 29]) the classification of the turbulence intensity of the wind is given by the following five thresholds: 10%, 15%, 20% and 30%. Figure 1 illustrates the wind observation for each Wednesday of the year 2017, during the 4 a.m. to 8 p.m. period, plotting the (10Hz frequency) wind speed, and in the same line the 40-minutes mean velocity components and the TI in (3.12). The TI is estimated using the mean $\langle q_t^{\text{obs}} \rangle$ computed over the time-scale $\zeta_{\text{TI}} = 10$-minutes and the reference mean velocity $\langle U_{(d)}^{\text{obs}} \rangle$ is computed on the entire period from 4 a.m. to 8 p.m. and updated for each considered days. As shown in Figure 1, the (arbitrary) selection of 46 Wednesday signals reveals a wide variety of behaviour. The lowest wind speed can be observed during November 8th, in contrast with December 27th having the highest wind speed during the afternoon. Regarding the values of the TI, the highest values were captured during June 21th and August 16th. We can observe also that, since the measurements were taken near the ground, the vertical component of the wind velocity stays closed to zero. Nevertheless, it adds a smoothing effect on the reconstruction of q_t^{obs} in (3.11).

3.5. Time depend regimes in the reduced model

The variety of magnitudes, and of jumpy behaviours illustrated in Figure 1 suggests some regimes dependency in the signal, easily explained by the dynamics of meteorological situations (wind direction changing the wall turbulence with the terrain, temperature, pressure, relative humidity conditions, modifying the stability of the ABL). In order to represent this regime dynamics in the uncertainty modelling, we allow the parameters to be period-depend, with periods limited in the presentation of this study to diurnal period (precisely from 4 a.m. to 8 p.m in local time) that we call 'day-period' arbitrary chosen along the annual cycle of the year 2017.

Formally, we are defining sub-periods, numbering from 1 to N_p and we allow the parameter values to be period dependent. This assumption is first made possible by the abundance of observations in each sub-period. Moreover Bayesian inference and the evaluation of the level of uncertainty in the calibration process allows us to assess it a posteriori (see Figure 5).

With this in mind, we can go further in the temporal dependence, in order to better capture the variation in the turbulent regimes represented by the parameter γ (see the orange curves of turbulence intensity in Figure 1). For the parameter γ only, and only during the posterior calibration Step one, we will refine each day-period, partitioning them in sub-signals of 20-minutes long, for a total of 48 sub-signals per day. Analysing the results of the time-dependence reduced model in Section 5, we will be able to show the relation between the γ-posterior calibration and the observed turbulent intensity statistic, and finally propose this last quantity as a predictive value for γ (see Relation 5.2 in Section 5).

4. Calibration and analysis of the reduced model

We now move forward to the next step: infer on the possible values of the parameters of the model (3.9), considering relevant observational dataset for wind energy application (see Section 3.4). To this aim, one could follow a frequentist approach by considering a single-value estimation for each parameters. Or one could follow a Bayesian approach, by assigning a probability to each parameter possible values. The Bayesian approach is often combined with Markov Chain Monte Carlo (MCMC) techniques: a Markov chain is constructed, sampling a distribution that converges with time to the stationary distribution of the parameters. Nevertheless, within Bayesian methods it is necessary to set a prior distribution for the parameters to then update this information using the observations.

In this section we construct a calibration procedure to infer the values of the parameters

$$C_\alpha = \frac{C_\varepsilon}{\kappa z(x_{\text{obs}})},$$

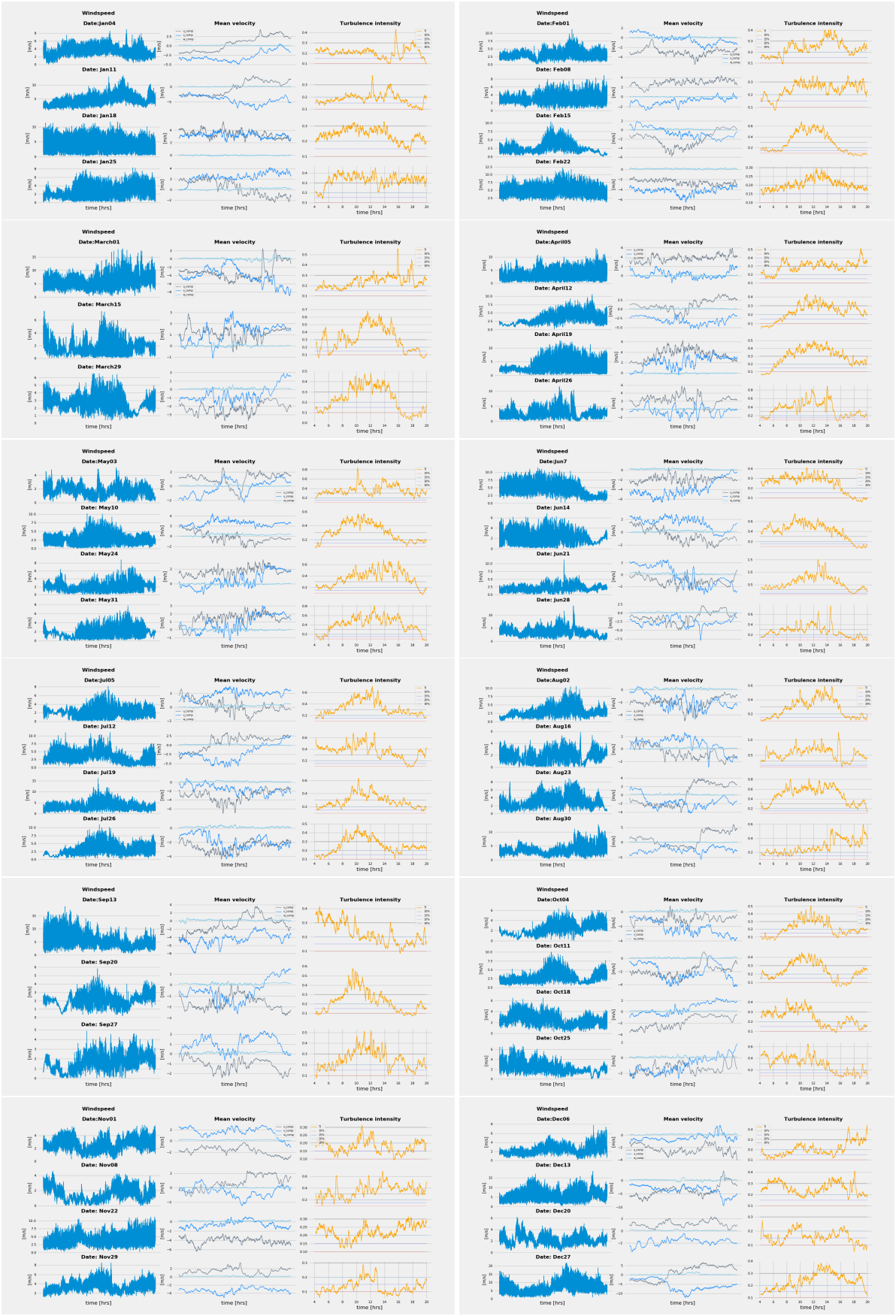


Figure 1: For each Wednesday in 2017, during the period of 4 a.m. to 8 p.m (local time), we plot the observed wind speed (in blue), the corresponding mean velocity (in blue scale) and turbulence intensity (in orange).

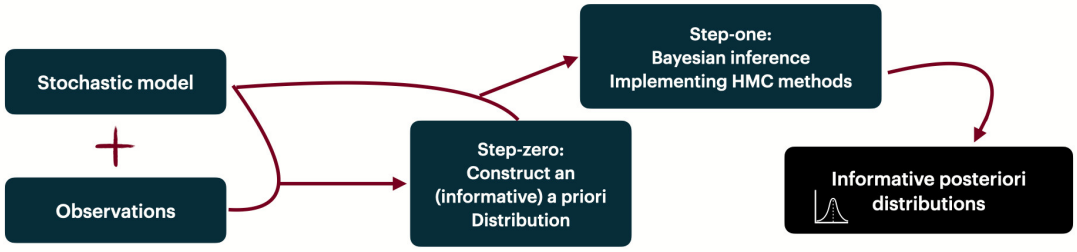


Figure 2: the proposed two-steps calibration procedure.

and γ. In order to simplify the analysis of the results, the Kolmogorov constant C_0 is considered as a prescribed constant equal to 1.9. Nevertheless, the Bayesian method used here can be extended in future work to include the calibration of C_0, which in view of Remark 2.1 may vary with cases. For the parameter C_α, a physically admissible prior distribution support is set using the typical values in the literature (see Remark 4.1 below). For the parameter γ however, we do not have a priori information besides an estimation of the support of its distribution. In order to provide a reliable calibration without additional parameters, we propose a two steps method summarized in Figure 2: in a Step zero, we construct the a priori distribution of the parameters through a learning step. In a Step one, we quantify the uncertainty of the parameters with Bayesian inference.

Before detailing the two steps of the calibration procedure, it is important to recall the nonlinear dependence of the coefficients of the CIR process (3.9) Θ, μ and σ in terms of the vector parameter (C_α, γ). Although the moments (3.10) and the transition density of the CIR process (3.9) are explicit in terms of Bessel functions, the tricky dependence of the parameters in these formulas complicates the computation of stable estimators by an optimisation problem. Then, for both steps of the calibration we introduce a discrete time approximation for the solution of the model (3.9). To this aim, we consider an homogeneous time step $\Delta t > 0$, some discrete times $t_n = n\Delta t$, and define the symmetrized Euler scheme [8] associated to the SDE (3.9) as:

$$\widehat{q}_{t_{n+1}} = |p_{t_{n+1}}|, \quad p_{t_{n+1}} = \widehat{q}_{t_n} + \Theta(C_\alpha,\gamma)\,(\mu(C_\alpha,\gamma) - \widehat{q}_{t_n})\,\Delta t + \sigma(\gamma)\sqrt{\widehat{q}_{t_n}}\,\left(W_{t_{n+1}} - W_{t_n}\right), \text{ for } n \leq N-1, \tag{4.1}$$

with initial condition $\widehat{q}_0 = q_0$, and N such that $N\Delta t$ equals the duration of the signal (scaled in seconds). Under the assumption $C_0 < 2$, the scheme (4.1) converges in law with a rate of order one [12, Theorem 2.3], i.e. for $f \in \mathcal{C}^4$ a bounded function with bounded derivatives, and sufficiently small Δt, there exists a constant C (independent on Δt) such that:

$$\sup_{n \leq N} |\mathbb{E}[f(q_{t_n}) - f(\widehat{q}_{t_n})]| \leq C(t_N)\,\Delta t.$$

Remark 4.1. *In order to set an interval of referential values for* $C_\alpha = \frac{C_\varepsilon}{\kappa z(x_{obs})}$, *we fix* $z(x_{obs}) = 30$ m. *We make varying the Von Kármán constant* $\kappa \in [0.287, 0.615]$ *[27]. We make also varying* $C_\varepsilon = C_\mu^{3/4}$, *with* $C_\mu \in [0.054, 0.135]$ *[27]. Therefore, we expect the values of* C_α *to be within the interval* $[0.0061,\ 0.0259]$.

The Kolmogorov constant is fixed to $C_0 = 1.9$, *in accordance with the values proposed within the literature pointed in Remark 2.1. With* $C_0 < 2$, *the condition ensuring the first order convergence of the time discretization scheme (4.1) is fulfilled.*

We now detail the calibration steps.

4.1. Step zero: Prior calibration

The maximum likelihood estimator (MLE) is a classical parameters calibration method for statistical models having an explicit density [18]. As already noticed in our case, due to the complexity of the CIR density in terms of the parameters, it is preferable (and common in such situation) to have the help of a discrete time approximation leading to a pseudo-likelihood computation. The symmetrized Euler scheme is far to be the only possible scheme for power-form coefficients SDEs –see e.g [13] and references therein, for other schemes and order rates, depending on C_0 and γ– but it has the advantage to be time-explicit, having the folded Gaussian law as transition density from $\widehat{q}_{t_k}$ to $\widehat{q}_{t_{k+1}}$. Moreover, knowing $\widehat{q}_{t_k}$, the probability $\mathbb{P}(\widehat{p}_{t_{n+1}} < 0)$ decreases with Δt [12, Lemma 3.7], allowing to approximate the folded Gaussian density on $[0, +\infty)$ for $\widehat{q}_{t_{k+1}}$, by the Gaussian one: for each $n = 0, \dots N-1$, we then assume in this a priori step the random variable $\widehat{q}_{t_{n+1}}$ to follow the distribution:

$$\widehat{q}_{t_{n+1}} \ \sim\ \mathcal{N}\big(\widehat{q}_{t_n} + \Theta(C_\alpha,\gamma)(\mu(C_\alpha,\gamma) - \widehat{q}_{t_n})\Delta t,\ \sigma^2(\gamma)\widehat{q}_{t_n}\Delta t\big). \tag{4.2}$$

From this, considering the compact set $D \subset \mathbb{R}^+ \times \mathbb{R}^+$ supporting the admissible values of the vector $\theta = (C_\alpha, \gamma)$, choosing Δt according to some data frequency, we compute the pseudo-maximum likelihood estimator as

$$\widehat{\theta} = \arg\max_{\theta \in D} \ \log p^{\theta}_{\Delta t}(q^{\text{obs}}_0, \dots, q^{\text{obs}}_{t_N}),$$

where, using the Markov property of solution of (4.1) and assuming $q^{\text{obs}}_{t_{n+1}}$ knowing $q^{\text{obs}}_{t_n}$ is distributed according to (4.2), the model density $p^{\theta}_{\Delta t}$ is expressed in terms of a product of Gaussian densities as follows

$$\begin{aligned}\widehat{\theta} &= \arg\max_{\theta \in D} \ \sum_{n=0}^{N-1} \log\Big(\frac{1}{\sqrt{2\pi\sigma^2(\gamma) q^{\text{obs}}_{t_n} \Delta t}} \exp\Big\{ - \frac{\left| q^{\text{obs}}_{t_{n+1}} - q^{\text{obs}}_{t_n} - \Theta(C_\alpha, \gamma)\left(\mu(C_\alpha, \gamma) - q^{\text{obs}}_{t_n}\right)\Delta t \right|^2}{2\sigma^2(\gamma) q^{\text{obs}}_{t_n} \Delta t} \Big\} \Big) \\ &= \arg\max_{\theta \in D} \ N\Big\{ -\frac{1}{2}\log\gamma - \frac{\widehat{M}_{2,-1}}{4C_0\Delta t\gamma} - \frac{C_R \widehat{M}_{1,0}}{2^{4/3} C_0} \left(\frac{C_\alpha}{\gamma}\right)^{2/3} - \frac{C_R^2 \Delta t \widehat{M}_{0,1}}{C_0 2^{8/3}} \left(\frac{C_\alpha}{\gamma}\right)^{1/3} C_\alpha \\ &\qquad + \frac{C_R^2 \Delta t}{C_0 2^{4/3}} \left(\frac{C_\alpha}{\gamma}\right)^{2/3} \gamma - \frac{C_R^2 \Delta t \widehat{M}_{0,-1}}{4C_0}\gamma \Big\},\end{aligned} \tag{4.3}$$

and where we use the notation: for any integers m_1 and m_2 equal to 0, 1 or 2.

$$\widehat{M}_{m_1,m_2} = \frac{1}{N} \sum_{n=0}^{N-1} (q^{\text{obs}}_{t_{n+1}} - |q^{\text{obs}}_{t_n}|)^{m_1} |q^{\text{obs}}_{t_n}|^{m_2}. \tag{4.4}$$

But here again – and contrary to the direct (Θ, μ, σ)–parameters pseudo-MLE calibration of CIR process that can be solved explicitly [55] – the optimal pair $(\widehat{C}_\alpha, \widehat{\gamma})$ is not easy to identify as our numerical tests leads to rather unstable approximation of an optimum. To avoid this computational issue, we appeal to the construction of an estimator for γ by using the formal convergence of the quadratic variation of a diffusion process as it was proposed in [39]. The main advantage in using quadratic variation estimator (QVE) is its simplicity from a computational point of view. In our case, the QVE for γ reads as:

$$\widehat{\gamma} = \frac{\widehat{M}_{2,0}}{2C_0\Delta t \widehat{M}_{0,1}}, \tag{4.5}$$

which is always positive and converges in probability, as N tends to infinity, to the parameter γ [39, Corollary 3.3].

Considering then the estimator $\widehat{\gamma}$ in (4.5), the optimisation problem (4.3) reduces to

$$\widehat{C}_\alpha = \arg\max_{c > c_*} \varphi(c),$$

with $\varphi(c) = (\widehat{\gamma} C_R \Delta t - \widehat{M}_{1,0})\, c^{2/3} - \frac{\widehat{\gamma}^{1/3} C_R \Delta t \widehat{M}_{0,1}}{2^{4/3}}\, c^{4/3}$, and $c_* > 0$ is a given lower bound including the values of C_α from expert knowledge (see Remark 4.1). Then, applying the first order optimality condition,

$$\widehat{C}_\alpha = \frac{\sqrt{2}}{\sqrt{\widehat{\gamma}}} \max\Big\{ c_* \frac{\sqrt{\widehat{\gamma}}}{\sqrt{2}}, \Big(\frac{\max\{\widehat{\gamma}\Delta t C_R - \widehat{M}_{1,0}\,, 0\}}{\widehat{M}_{0,1} \Delta t C_R} \Big)^{3/2} \Big\}. \tag{4.6}$$

Furthermore, since $\widehat{\gamma}\Delta t C_R - \widehat{M}_{1,0} > 0$, we have $\varphi''(c) < 0$, for all $c > 0$. Then, the second order optimality condition allows to conclude that the maximum of φ is achieved with (4.6).

Finally, we get our Step zero estimators (4.6) and (4.5). From Identity (4.5), the condition $\widehat{\gamma}\Delta t C_R - \widehat{M}_{1,0} > 0$ becomes $C_R \widehat{M}_{2,0} - 2C_0 \widehat{M}_{1,0} \widehat{M}_{0,1} > 0$. This condition has been checked to be always satisfied for our wind time series described in Sections 3.4. A summary of the results produced with this first calibration step is presented in Section 5.

4.1.1. Time dependent regimes and prior calibration

To go further on the quantification of uncertainty for the model parameters, one has to propose an a priori distribution which is both sufficiently informative and possibly admits values coming from expert knowledge. To this aim, for each day-period of interest d in a selection $\mathcal{S}$, we first compute the prior estimators $\widehat{\gamma}(d)$ and $\widehat{C}_\alpha(d)$

given, respectively, by Equations (4.5) and (4.6) from the time series corresponding to segment d. From this, we obtain a family of estimators

$$\Sigma := \{(\widehat{C}_\alpha(d),\ \widehat{\gamma}(d)),\ \text{for } d \text{ in the selection of day-periods } \mathcal{S} \text{ in the year 2017}\},$$

from which we define the truncated Gaussian a priori distributions:

$$\gamma \sim \mathcal{N}^+\left(\overline{\Gamma}(\mathcal{S}), \mathbb{V}_\Gamma(\mathcal{S})\right), \qquad C_\alpha \sim \mathcal{N}^+\left(\overline{C}(\mathcal{S}), \mathbb{V}_C(\mathcal{S})\right), \tag{4.7}$$

where $\overline{\Gamma}(\mathcal{S}) = \frac{1}{\#\mathcal{S}}\sum_{d\in\mathcal{S}}\widehat{\gamma}(d)$ and $\overline{C}(\mathcal{S}) = \frac{1}{\#\mathcal{S}}\sum_{d\in\mathcal{S}}\widehat{C}_\alpha(d)$ denote the empirical means of the $\widehat{\gamma}$ and $\widehat{C}_\alpha$ estimates, and $\mathbb{V}_\Gamma$ and $\mathbb{V}_C$ their corresponding empirical variances over the day-periods in $\mathcal{S}$.

4.2. Step one: Posterior calibration

Inspired by the work of Edeling et al. [27], we improve the estimations computed in Step zero by quantifying the uncertainty on the two parameters distribution of the reduced model (3.9). With the help of MCMC method, we apply a Bayesian inference to update our initial guess in (4.7) on the distribution of the parameters in the prior calibration step.

4.2.1. Statistical model for the uncertain parameters

Assume that the observed TKE on the interval $[0, T]$ are independent and identically distributed observable vector random variables $\boldsymbol{q}^{\text{obs}} \in (0, +\infty)^{N+1}$, indexed with $\{0 = t_0, t_1, \ldots, t_N = T\}$. As the name suggests, Bayesian inference uses Bayes' Theorem in order to extract information from the observed values and infer a more realistic distribution for the parameters. Denoting by $\pi(\theta|\boldsymbol{q}^{\text{obs}})$ the probability density of the parameter vector $\theta = (\gamma, C_\alpha)$ given the realisation $\boldsymbol{q}^{\text{obs}}$, by $p(\cdot|\theta)$ the probability density of the model given the parameters (likelihood function), and by p_θ the prior distribution of θ in (4.7), we get from Bayes' Theorem that

$$\pi(\theta|\boldsymbol{q}^{\text{obs}}) = \frac{p(\boldsymbol{q}^{\text{obs}}|\theta)\ p_\theta(\theta)}{p(\boldsymbol{q}^{\text{obs}})}, \tag{4.8}$$

where $p(\boldsymbol{q}^{\text{obs}})$ is the distribution of the observed data. For this latter distribution, we set a statistical model through our model for instantaneous TKE which comes with observational error $\boldsymbol{\mathcal{E}}$ as:

$$\boldsymbol{q}^{\text{obs}}(\theta) = \widehat{\boldsymbol{q}}(\theta) + \boldsymbol{\mathcal{E}}, \tag{4.9}$$

where the independent random vector variable $\widehat{\boldsymbol{q}}(\theta)$ is identically distributed with the equilibrium law (for a given θ) of the discrete-time process (4.1); the random vector $\boldsymbol{\mathcal{E}}$ is assumed following the logistic distribution with zero mean and scale parameter σ to be estimated from the data. The choice of the logistic distribution is made from a preliminary analysis comparing the histogram of errors with a set of explicit mean-variance parametrized probability densities (illustrated in Figure 4 of Section 5).

Metropolis-Hastings (MH) Algorithm. A widely used method to generate samples from the posterior distribution is the Metropolis-Hasting algorithm. In our case, this iterative algorithm, sampling for the Markov chain (θ_n, n), takes the following form: start from an initial value θ_0. At the nth-iteration, from θ_n to θ_{n+1}, proceed as follow:

1. Simulate $\widetilde{\theta} \sim \rho(\widetilde{\theta}|\theta_n)$, and $u \sim \textit{Uniform}(0, 1)$, where ρ is a proposed transition density.
2. Compute

$$a := \min\left\{1, \frac{p_\theta(\widetilde{\theta})p(\boldsymbol{q}^{\text{obs}}|\widetilde{\theta})\rho(\theta_n|\widetilde{\theta})}{p_\theta(\theta_n)p(\boldsymbol{q}^{\text{obs}}|\theta_n)\rho(\widetilde{\theta}|\theta_n)}\right\}. \tag{4.10}$$

3. If $u < a$, the simulated state is accepted and $\theta_{n+1} = \widetilde{\theta}$. Else, the state $\widetilde{\theta}$ is rejected and we keep the previous state, i.e. $\theta_{n+1} = \theta_n$.

Notably, the second step of the algorithm maximizes the posterior of the parameters with their transition, given the observed data. The choice of a prior distribution close to the posterior one improves the convergence of the chain to its stationary distribution $\pi(\cdot|\boldsymbol{q}^{\text{obs}})$. Nevertheless, a common issue in the implementation of this method is the computational time needed to explore properly the state space in high dimension. Through the sampling process, we not just infer the values of the parameters θ but also the statistical model (4.9).

Hamiltonian Monte Carlo (HMC) is an alternative method to efficiently explore the state space of the MH framework. HMC uses the well-known Hamiltonian system to mimic the dynamics of a particle having the logarithm of the target posterior distribution (which must be smooth enough) as potential energy. To this aim, the HMC method introduces a synthetic variable ψ, (usually sampled as a Gaussian random variable), describing the momentum. Assuming the conservation of the total energy, we consider the Hamiltonian function as the sum of kinetic energy K (fixed to $\frac{1}{2}\|\psi\|^2$ for instance), and potential energy (as the log posterior density)

$$H(\theta, \psi) = K(\psi) - \sum_{q \in \boldsymbol{q}^{\text{obs}}} \log p(q|\theta) - \log p_\theta(\theta), \tag{4.11}$$

and the parameter vector θ represents now the position of a particle following the Hamiltonian equations:

$$\begin{cases} \dfrac{d\theta}{dt} = \dfrac{\partial H}{\partial \phi}, \\ \dfrac{d\psi}{dt} = -\dfrac{\partial H}{\partial \theta}. \end{cases} \tag{4.12}$$

The exchange between kinetic and potential energies during the discrete time simulation of the Hamiltonian dynamics ensures that the evolution of the particle generates the contours of the target distribution π, where the latter is obtained by computing the marginal distribution of the joint position-momentum posterior. Among the main advantages within HMC methods, we underline the preservation of the volume, the efficient exploration of the space, the reversibility of the dynamics (which is crucial to ensure that the MCMC updates preserves the target distribution). These features allow HMC algorithms to converge to high-dimensional target distributions much more quickly than simpler methods such as random walk Metropolis or Gibbs sampling. Nonetheless, the introduction of this auxiliary momentum duplicates the number of variables, and therefore the computational cost.

Sampling the momentum from the conditional distribution (assuming that the current point θ_n is in the contour of the target distribution) we generate trajectories moving -in time- through the entire phase space while being constrained to the typical set. Thus, to propose a new state θ_{n+1} for the Markov chain, the trajectory is projected back to the parameter space, and finally the state is accepted/rejected following a decision criterion similar to the one implemented in the Metropolis-Hasting algorithm. For a more detailed introduction to HMC methods we refer the interested reader to the work of Neal [43].

Since usually the likelihood function involved in (4.11) is not explicit, and even so, the solution to Hamiltonian system is not explicit as well, the HMC method requires a discrete-time approximation to Equations (4.12). The quality of numerical approximation relies on the ability of the simulated trajectories to not drift away from the exact energy level set. One common approach is the leapfrog scheme which start with half step for the momentum variables, then do a full step for the position using the updated momentum, and finally complete the remaining half step for the momentum. The main drawback of solving (by approximation) the dynamics (4.12) is the need to specify the terminal time and the size of the steps in the trajectory. Among the methods implemented for the sampling of Monte Carlo Markov Chains with Hamiltonian method, we use the No U-Turn sampler [28] that allows automatic tuning of the step size and number of simulation steps.

4.2.2. Algorithm for the Bayesian calibration

We now detail the entire Bayesian calibration algorithm and discuss its specificities for the model (3.9).

Consider an arbitrary selection of day-periods $\mathcal{S}$. For d in $\mathcal{S}$, we proceed as follows for the Bayesian calibration of $\widehat{\theta}(d) = (\widehat{C}_\alpha(d), \widehat{\gamma}(d))$.

Selection of frequencies. We recall that for the Bayesian algorithm, we need independent realisations of variables $\boldsymbol{q}^{\text{obs}}$ selected from the observed time series. To do so, we compute the autocorrelation of the signal and select a compromise between independence and the length of the sample: we therefore choose a frequency $\xi_{C_\alpha} = 1/30\ \text{s}^{-1}$ for the Bayesian calibration of C_α on the 16 hours duration of the day-period d, and a frequency of $\xi_\gamma = 1/5\ \text{s}^{-1}$ for the Bayesian calibration of γ for each sub periods of 20 minutes of the day-period.

For the calibration coherency between Step zero and Step one, in Step zero, we stay with the frequency $\xi_{C_\alpha} = \xi_\gamma = 1/30\ \text{s}^{-1}$, calibrating with QVE and QMLE the parameter pair $(\widehat{C}_\alpha(d), \widehat{\gamma}(d))$, assumed constant on the entire day-period. These chosen frequencies set a unique time step Δt to use in (4.2). Moreover, due to the high frequency of jumps in the observations, this rather low frequency helps to bring out the diffusive behaviour in the observations.

For convenience we denote $[0, T]$ the day-period d, with dates $\{0 = t_0, t_1, \ldots, t_N = T\}$, and proceed as follows:

0. *Compute the prior distribution* (4.7).

1.1 *Estimate the observational error.* We estimate the scale parameter $\widehat{\sigma}(d)$ of the logistic distribution appearing in the statistical model retained for the observational error $\boldsymbol{\mathcal{E}}$ in (4.9). For this, we first compute the empirical mean

$$\mathbb{E}[\boldsymbol{\mathcal{E}}(d)] \simeq \frac{1}{N}\sum_{n=1}^{N} q_{t_n}^{\text{obs}} - \widehat{q}_{t_n}(\widehat{\theta}(d)),$$

simulating $\widehat{q}$ using (4.1) with $\Delta t = \frac{1}{\xi_{C_\alpha}} = 30$ s (the lower frequency is used here as we need an independent sample to approximate the mean) and with $\widehat{\theta}(d) = (\widehat{C}_\alpha(d), \widehat{\gamma}(d))$ computed in Step zero. Computing the empirical variance with the same sample, we obtain the approximate scale $\widehat{\sigma}^2(d) = \frac{3}{\pi^2}\mathbb{V}[\boldsymbol{\mathcal{E}}(d)]$.

1.2 *Bayesian calibration of* γ. As announced in Section 3.5, γ is allowed to be estimated on refined sub-signals along a day-period $[0,T] = \bigcup_{0\le i\le S}[T_i, T_{i+1}]$, with $T_{i+1} - T_i = 20$-minutes long. In this step, the parameter $\widehat{C}_\alpha(d)$ is fixed, given by the Step zero estimator. For each $i \in \{1, \dots, S = 48\}$, we estimate the density of $\gamma(d,i)$ considering the sub-signal $\boldsymbol{q}^{\text{obs}}\big|_{[T_i,T_{i+1}]}$ with the higher frequency ξ_γ and the time step $\Delta t = \frac{1}{\xi_\gamma} = 5$ s for the simulation of $\widehat{\boldsymbol{q}}_t$. Starting from the prior distribution $\gamma(d,i) \sim \mathcal{N}^+(\overline{\Gamma}, \mathbb{V}_\Gamma)$ in (4.7), and observation error $\mathcal{E} \sim \textit{Logistic}(0,\sigma)$ with scale $\sigma \sim \textit{Lognormal}(\widehat{\sigma}(d), 1)$, we apply the HMC method in order to sample the posterior distribution of $\gamma(d,i)$ with the statistical model

$$\boldsymbol{q}^{\text{obs}}\big|_{[T_i,T_{i+1}]} = \widehat{\boldsymbol{q}}(\gamma(d,i), \widehat{C}_\alpha(d)) + \boldsymbol{\mathcal{E}}.$$

In particular, by computing $\mathbb{E}[\gamma(d,i)]$ from the empirical mean of the HMC sampling of size 2000, we obtain a piecewise constant mean production term process

$$\bar{\gamma}_t = \sum_{i=0}^{S} \mathbb{E}[\gamma(d,i)]\mathbb{1}_{[T_i,T_{i+1}]}(t). \tag{4.13}$$

1.3 *Bayesian calibration of* C_α. Starting from the prior distribution $\mathcal{N}^+\left(\overline{C}(\mathcal{S}), \mathbb{V}_C(\mathcal{S})\right)$ in (4.7), and observation error $\boldsymbol{\mathcal{E}}$ as in 1.2, we apply a HMC method in order to sample the posterior distribution of $C_\alpha(d)$ with the statistical model

$$\boldsymbol{q}_t^{\text{obs}}|_{[0,T]} = \widehat{\boldsymbol{q}}_t(\bar{\gamma}_t, C_\alpha(d)) + \mathcal{E}_t.$$

In this step we consider $\boldsymbol{q}^{\text{obs}}$ with frequency ξ_{C_α} and $\Delta t = \frac{1}{\xi_{C_\alpha}}$.

For the HMC implementation, we have used the Python package PyMC3 [51] with NUTS step method.

Convergence Diagnostic Tests. To check the convergence of the HMC method to the posterior distribution, several convergence diagnostics can be assessed. For instance, we can verify if the Markov chain explores the state space thoroughly, check the convergence of the empirical mean or analyse if the simulated values are uncorrelated. Along with ad hoc convergence diagnostics, formal statistical diagnostics have been implemented. The convergence of each of the Markov chains constructed in steps 1.i has been tested through the following:

- *Geweke diagnostic* that analyses the similarity (mean and variance) of segments from the beginning and the end of a single chain, and check if the samples are drawn from the stationary distribution.
- *Gelman-Rubin diagnostic* (also called $\widehat{R}$-hat test) that uses multiple chains to check the lack of convergence by computing the ratio between the within-chain variance and the posterior variance estimate for the pooled traces. This quotient converges to 1 when each of the traces is a sample from the target posterior.

5. Calibration results and model evaluation against data

First, we recall that for the presentation of the results summarized in this section, we have chosen to show a selection set $\mathcal{S}$ arbitrary day-periods corresponding to wind measurements observed every Wednesday (46 day-periods) of the year 2017, starting from 4 a.m. to 8 p.m. (local time), maintaining the representativeness of the different regimes and seasonality (see Figure 1).

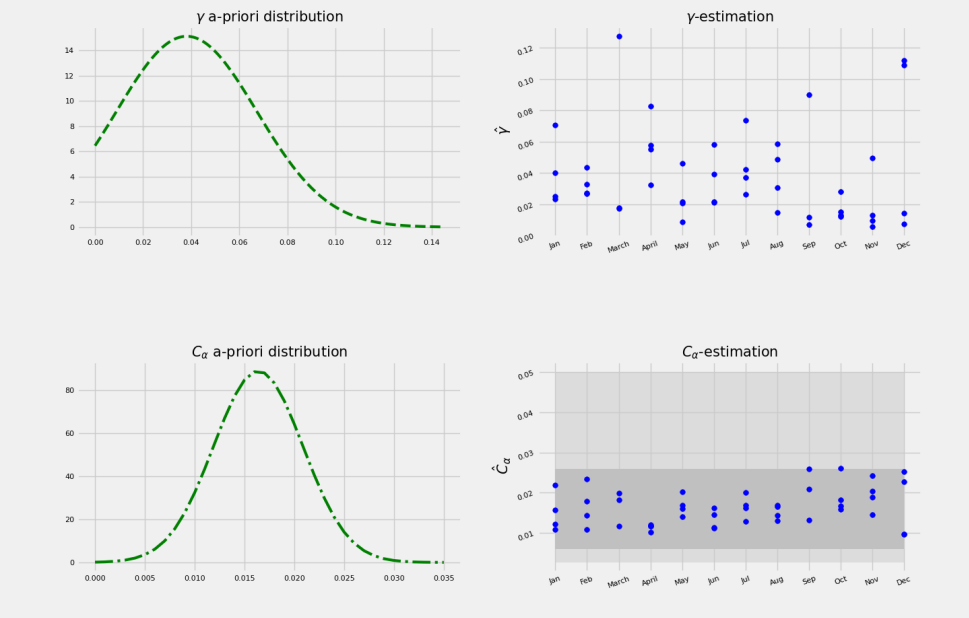


Figure 3: Prior calibration results: point-estimations (right figures) for each Wednesday of 2017, and resulting prior distribution (left figures) from (4.7). The dark grey area in the representation of the C_α calibration corresponds to the reference interval given in Remark 4.1.

In the following, we illustrate the results of the calibration, describing the key findings of each step. Before calibrating with the observations, we have tested each step using synthetic data generated from the time-discretisation of the SDE (3.9) used with fixed parameters, from which we recovered the input values (results not shown here).

In Step zero, after computing the point-estimates $(\widehat{C}_\alpha(d), \widehat{\gamma}(d), d \in \mathcal{S})$ with (4.5) and (4.6), we construct the prior distribution of the parameter vector as described in (4.7). The results of this step are illustrated in Figure 3, with the collection of point-estimations (right) and the informative a priori distribution (left) for each parameter. Concerning $\widehat{C}_\alpha$ point-estimations, it should be stressed that all the obtained estimates fall within the reference range of C_α values given in the literature and reported in Remark 4.1, without the need to impose the minimum value c_* in (4.6), which glimpses the accuracy of the model. Concerning the prior calibration of γ, the $\widehat{\gamma}(d)$ estimates are all positive, spread in a relatively larger interval (and with a large variance), which is coherent with the idea of decomposing the calibration of γ on sub-signals.

In order to validate the estimates obtained from Step zero, we consider the theoretical moment $\mathbb{E}[q_\infty]$ given in (3.8) and, along each day, we compare the quotient $(\frac{\sqrt{2}\widehat{\gamma}(d)}{\widehat{C}_\alpha(d)})^{2/3}$ against the time-averaging $\frac{1}{\#\text{obs}} \sum_t q_t^{\text{obs}}$. Table 1 summarizes these values showing the accuracy of the prior step of the calibration procedure.

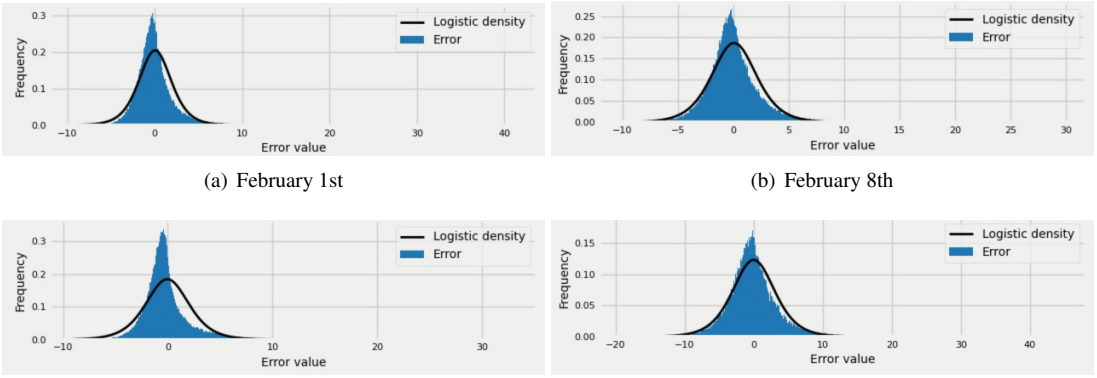
Concerning the calibration of the distribution of the observation error $\boldsymbol{\mathcal{E}}$ in (4.9), we mention that after several numerical tests, we have chosen a logistic distribution with scale parameter $\widehat{\sigma}^2(d) = \frac{3}{\pi^2}\mathbb{V}[\boldsymbol{\mathcal{E}}(d)]$ approximated with a Monte Carlo method. In Figure 4, we plot the empirical distribution of $\boldsymbol{\mathcal{E}}(d)$ resulting from (4.9), for $d =$ February 1st, 8th, 15th, 22th. The black curves plotted in the figure represent the logistic densities adjusted with the moments of the sample. Here, we can see that some days adapt better to the density (for example February 22th in contrast to February 15th). However, given the thinness of the empirical density tails, in general the logistic distribution is the simplest distribution that best fits.

In Figure 5, we summarize the results obtained for the Bayesian calibration of C_α, with the posterior mean estimations reported in Figure 5(b) for each selected day. The posterior mean estimates have been computed as the average on the Markov chain sample $\{C_{\alpha,n}(d), n \geq 1\}$ that passes convergence diagnostic tests according to the chosen length of the chain. So we can consider these empirical means converging towards the true parameter value. The box plots reported in Figure 5(a) have been constructed from the combination of all the Markov chain samples in a given month, by computing the quartiles of each set. Then, for each month, the range of the box comprises values between 25% and 75% of the inferred values for C_α during that month. Additionally, Figure 5(c) illustrates the exploration (right) of the state space for two samples (red and blue) along the Markov chain, for the particular day of February 15th, with their associated posterior distribution (left), having convergence $\widehat{R}$-hat statistic equals 1.

We emphasize that the estimated C_α values fall again within the reference interval $[0.0061, 0.0259]$, this time

Month	**Quotient** $(\frac{\sqrt{2}\widehat{\gamma}(d)}{\widehat{C}_\alpha(d)})^{2/3}$	**Time-average** $\frac{1}{\#\mathbf{obs}}\sum_t \boldsymbol{q}_t^{\mathbf{obs}}$	**Absolute error**
	1.32326987	1.32352805	1.95e-04
January	2.34663793	2.34799496	5.78e-04
	4.0458171	4.047094	3.16e-04
	2.19174731	2.1923935	2.95e-04
	1.64077689	1.64175207	5.94e-04
February	1.92520322	1.92621081	5.23e-04
	1.5771361	1.57724267	6.76e-05
	3.18971747	3.18980458	2.73e-05
	6.14997814	6.15392597	6.42e-04
March	1.17901587	1.17918501	1.43e-04
	1.22671331	1.22699872	2.33e-04
	3.99699439	3.99936725	5.93e-04
April	2.43719514	2.43752582	1.36e-04
	4.6508576	4.65245935	3.44e-04
	3.5330159	3.53318812	4.87e-05
	0.80623555	0.80670673	5.84e-04
May	2.18575829	2.18624388	2.22e-04
	1.53782389	1.53811257	1.88e-04
	1.63677343	1.63688688	6.93e-05
	3.76661667	3.76674479	3.40e-05
June	1.51005644	1.51051385	3.04e-04
	1.65372866	1.65381007	4.92e-05
	2.88200534	2.88234525	1.18e-04
	1.51360975	1.51388343	1.81e-04
July	2.55681663	2.55779937	3.84e-04
	3.36248203	3.3629028	1.25e-04
	2.39722227	2.39762117	1.66e-04
	2.59128586	2.59280967	5.88e-04
August	1.29836408	1.29888482	4.01e-04
	2.22652455	2.22708633	2.52e-04
	2.88813526	2.88946423	4.60e-04
	4.53006814	4.53197261	4.20e-04
September	0.74429092	0.74442679	1.83e-04
	0.62351139	0.62381067	4.80e-04
	1.12645897	1.12662812	1.50e-04
October	1.78036412	1.78109514	4.10e-04
	0.77009635	0.77025073	2.00e-04
	1.11176981	1.11209757	2.95e-04
	0.68282708	0.68306123	3.43e-04
November	0.53527354	0.53539273	2.23e-04
	2.85532726	2.8558594	1.86e-04
	0.98429323	0.98442425	1.33e-04
	0.9223277	0.92314383	8.84e-04
December	6.28067272	6.28366521	4.76e-04
	0.55240658	0.55471823	4.17e-03
	6.44366922	6.44440406	1.14e-04

Table 1: Validation of the prior calibration through Identity (3.8).



(a) February 1st (b) February 8th

(c) February 15th (d) February 22th

Figure 4: Observation error distribution, introduced in Subsection 4.2.1, approximated with Monte Carlo method. Histogram (blue bars) are compared with adjusted logistic density (in black).

the average values being closer to 0.01, except during October, November and December, months for which we can observe more variability and lower turbulence intensity.

Regarding the calibration of the parameter γ, in Figure 6 we summarize the results of the calibration of each sub-signal during February, 2017. More precisely, in Figure 6(a) we show the box-plot for each 20 minutes-length sub-signal, where we notice that the variance of those parameters that reach low values is quite small, suggesting that the state 0 is a kind of attractor point. On the other hand, Figure 6(b) shows the comparison between the mean posterior estimations (associated with each sub-signal), the mean γ during the 16 hours period for each day (coloured solid lines) and the prior estimator computed with quadratic variation estimator. From the results of the calibration of γ we highlight that, despite the fact that the training step estimator and the (daily) average estimator are quite close, the variation of γ is significant. The calibration by segments represents an undoubted improvement taking into account the time dependence of the production coefficient and the subsequent ability of the model to adapt to the data.

Once the calibrations in Steps zero and one are performed, we would like to verify the ability of the model in replicating the observations. To do so, we compute the 95% confidence interval for the trajectories $(\widehat{q}_{t_n}, n)$ and compare this confidence interval against the observations. Precisely, we use the inferred values $\widehat{C}_\alpha(d)$ illustrated in Figure 5(b) and the time dependent mean $\bar{\gamma}(t) = \sum_{i=0}^{S} \mathbb{E}[\gamma(d,i)]\mathbb{1}_{[T_i,T_{i+1}]}(t)$, for each calibrated day d, to construct the associated discrete time instantaneous turbulent kinetic energy with the symmetrized Euler scheme:

$$\begin{aligned} \widehat{q}_{t_{n+1}} &= |p_{t_{n+1}}|, \\ p_{t_{n+1}} &= \widehat{q}_{t_n} + C_R\bar{\gamma}_t\Delta t - C_R\left(\frac{\widehat{C}_\alpha^2(d)\bar{\gamma}_t}{2}\right)^{1/3}\widehat{q}_{t_n}\Delta t + \sqrt{2C_0\bar{\gamma}_t}\sqrt{\widehat{q}_{t_n}}(W_{t_{t+1}} - W_{t_n}),\ \widehat{q}_0 = \boldsymbol{q}_0^{\text{obs}}, \end{aligned} \tag{5.1}$$

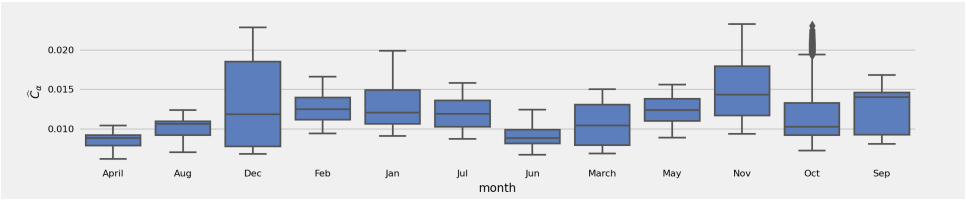
where $(W_{t_{t+1}} - W_{t_n}) \sim \mathcal{N}(0, \Delta t)$, with the time step Δt chosen according to the desired data frequency of $1/30\ \text{s}^{-1}$. Time by time, we estimate the 95% confidence interval of the trajectories $(\widehat{q}_{t_n}, n)$. In Figure 7, we observe that the confidence interval (in black) tightly envelops the observed trajectory (coloured) for $d =$ February 1st, 8th, 15th, 22th, validating the calibration process and the probabilistic model (3.9). Similar results were obtained for the whole data set.

5.1. The 10-minutes turbulence intensity as a substitute to the calibration of $(\bar{\gamma}_t)$

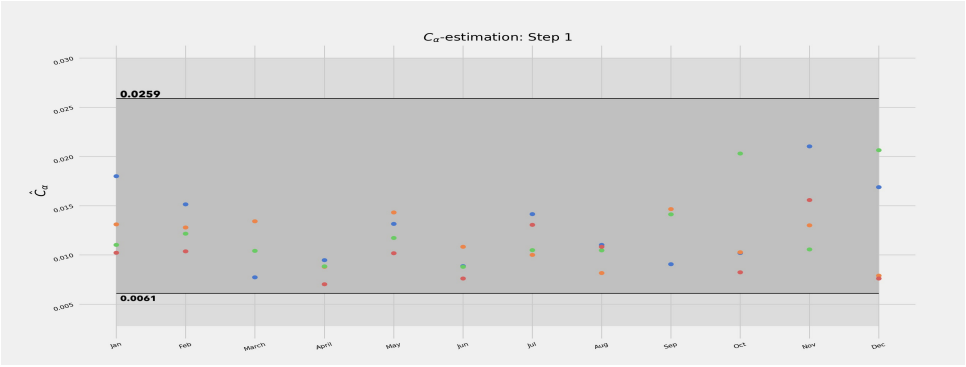
When comparing the curves of the time evolution $t \mapsto \bar{\gamma}_t$ in Figure 6(b) with those of the turbulent intensity $t \mapsto I_t$, defined in Equation (3.12) and illustrated in Figure 1, for the same days we can observe a good similarity of their behaviours even if the magnitudes of the two curves differ.

Let us going further into the connection between the turbulence intensity I_t and the value of the production coefficient γ_t. Assuming the stationary regime, and considering Equation (3.9), we would expect the instantaneous kinetic energy to oscillate around the mean. Then, if we apply the expectation operator on both sides of (3.9), and consider –by argument of stationarity – the zero value for the left-hand side, we get that for all $t \geq 0$

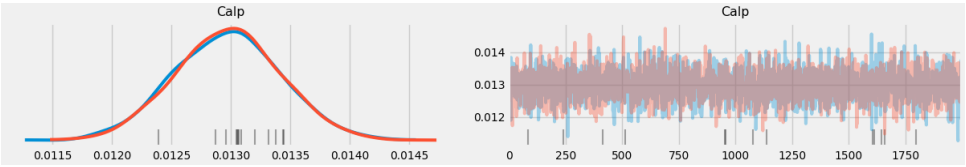
$$\gamma_t - \left(\frac{C_\alpha^2\gamma_t}{2}\right)^{1/3}\mathbb{E}[q_t] = 0,$$



(a) Box plots within month of posterior distribution of $\widehat{C}_\alpha$.

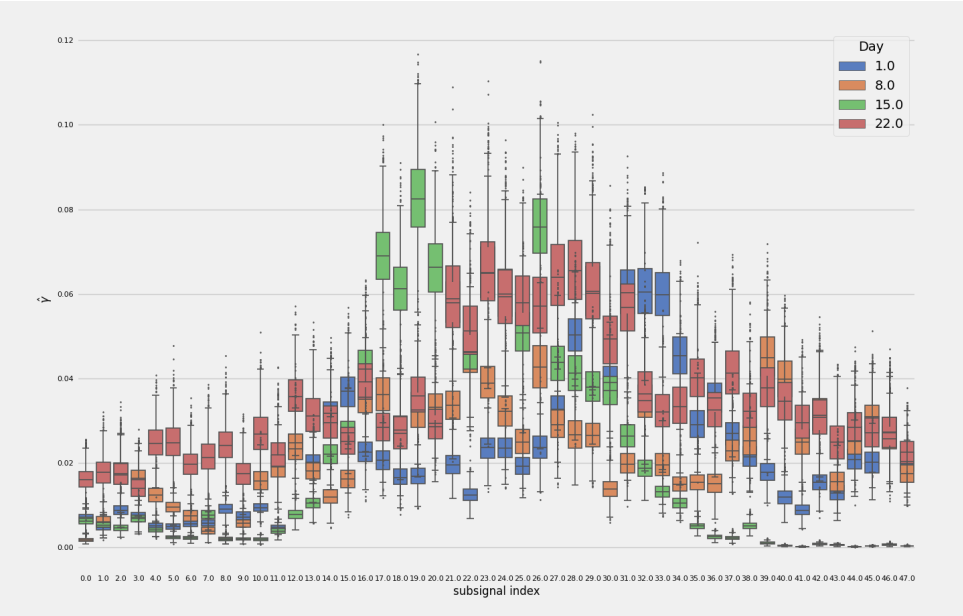


(b) Posterior mean estimations of $\widehat{C}_\alpha(d)$.

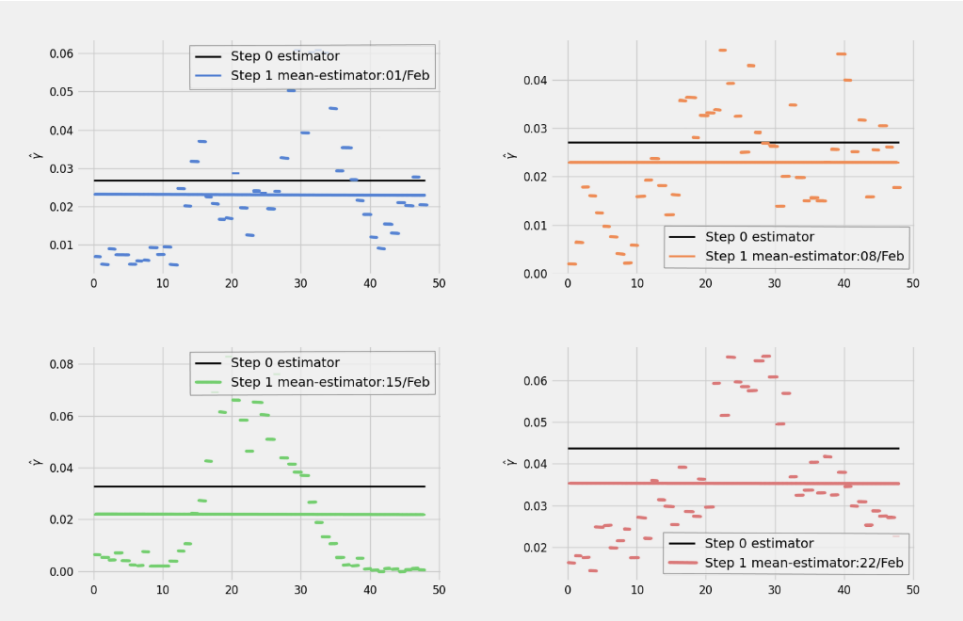


(c) Exploration of the state space and posterior distribution.

Figure 5: Bayesian calibration of C_α for all Wednesdays of 2017. In (a) the box plots constructed from the Markov chains samples in a given month. In (b) the obtained posterior mean estimations for each day. The colors blue, orange, green, red correspond respectively to occurrence order of the Wednesdays in each month. In (c), two examples of exploration of the state space with the Markov chain are illustrated (right), with the corresponding posterior distribution of C_α (left).

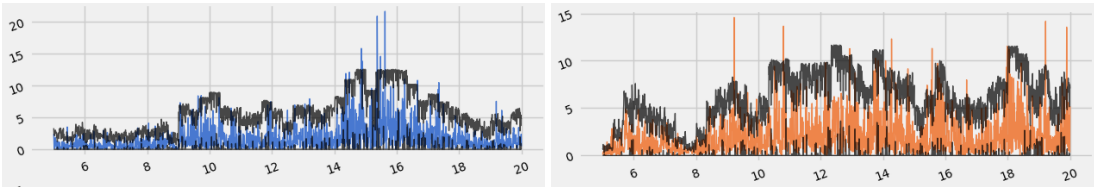


(a) Box plot of the $\gamma(d, i)$ for $i = 1, \ldots, 48$.



(b) The dynamics $t \mapsto \bar{\gamma}_t$ obtained from Step one for the same four days; the horizontal lines are the level of the means over the period (the black line is the Step zero estimator).

Figure 6: Bayesian calibration of γ during February, 2017: box plot for each 20 minutes-length sub-signal (6(a)) and comparison between Step zero and Step one with mean posterior estimations (6(b)) (small line-segments) day-mean posterior estimator (solid color line) and prior estimator (black line). The colors blue, orange, green, red correspond respectively to occurrence order of the Wednesdays in the month.



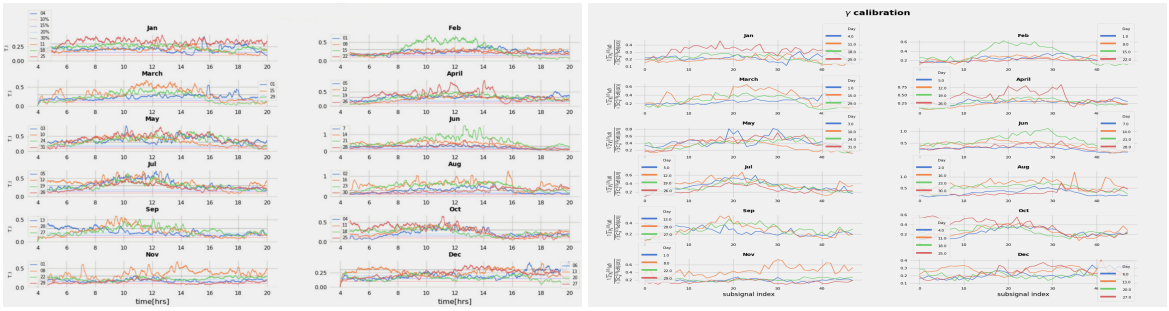
(a) $\boldsymbol{q}^{\text{obs}}$ for $d =$ Feb 1st

(b) $\boldsymbol{q}^{\text{obs}}$ for $d =$ Feb 8th

(c) $\boldsymbol{q}^{\text{obs}}$ for $d =$ Feb 15th

(d) $\boldsymbol{q}^{\text{obs}}$ for $d =$ Feb 22th

Figure 7: Validation of the calibration procedure: instantaneous turbulent kinetic energy observed during February, 2017, between 5 a.m and 8 p.m (color plots) using the frequency of 1/30 s^{-1}, and confidence interval (plotted in black) of $(\widehat{q}_{t_n}, n)$ in (5.1) using the posterior mean values $\widehat{C}_\alpha(d)$ and the time dependent mean $\bar{\gamma}(t) = \sum_{i=0}^{S} \mathbb{E}[\gamma(d,i)]\mathbb{1}_{[T_i,T_{i+1}]}(t)$, $\Delta t = 30$ s.



(a) Turbulence intensity I_t.

(b) Transformation $\frac{(\sqrt{2}\bar{\gamma}_t)^{1/3}(d)}{\sqrt{3}\widehat{C}_\alpha^{1/3}(d)\|\langle U_{(d)}^{\text{obs}}\rangle\|}$.

Figure 8: Validation of Identification (5.2).

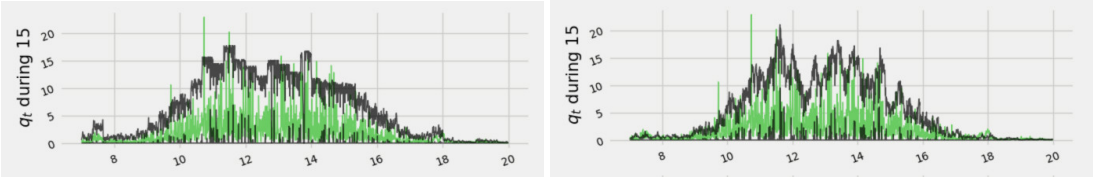
from which we deduce with (3.12) the formal relation:

$$\gamma_t = \frac{C_\alpha}{\sqrt{2}}\left(\sqrt{3}\|\langle U_{(d)}^{\text{obs}}\rangle\| I_t\right)^3. \tag{5.2}$$

In Figure 8 we verify the validity of the relation (5.2) by comparing the observed turbulence intensity I_t against the transformation $\frac{(\sqrt{2}\bar{\gamma}_t)^{1/3}}{\sqrt{3}\widehat{C}_\alpha^{1/3}(d)\|\langle U_{(d)}^{\text{obs}}\rangle\|}$ for each calibrated day. Then, assuming the stationary regime, the unquestionable similarity between TI and the transformation of $\bar{\gamma}_t$ from (5.2) suggests to consider directly $\bar{\gamma}_t$ in the model to be given dynamically by the 10-minutes turbulence intensity (3.12). This observation opens to a simpler and much more efficient calibration procedure for wind forecasting or near-casting applications.

Following the previous procedure leading to Figure 7 for the validation of the calibration method, we generate confidence intervals (with 95% confidence level) for the solution of the model (3.9) using the numerical scheme (5.1), but with input $\bar{\gamma}_t$ given by relation (5.2). An updated information on the turbulence intensity is needed at each time sub-interval. The parameter C_α was sampled from the distribution $\mathcal{N}(1.18 \times 10^{-2}, 1.21 \times 10^{-5})$ obtained from its within-year posterior distribution. In Figures 9 and 10 we illustrate the ability of this calibration method to predict the instantaneous turbulent kinetic energy. Figure 9 shows the impact of using the turbulent intensity statistic to predict $\bar{\gamma}_t$ combined with the sampling of the posterior C_α distribution, comparing with the Bayesian calibration procedure result in Figure 7(c). Figure 10 shows the predicted 95% confidence interval for two days chosen randomly in the year 2017.

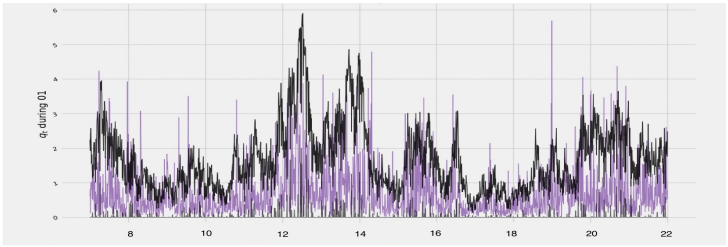
We end this section with a more global in time evaluation of the model, by comparing the distribution of the turbulent wind speed $\|u'\|$ obtained over a year from the TKE model with the one from the observations. More



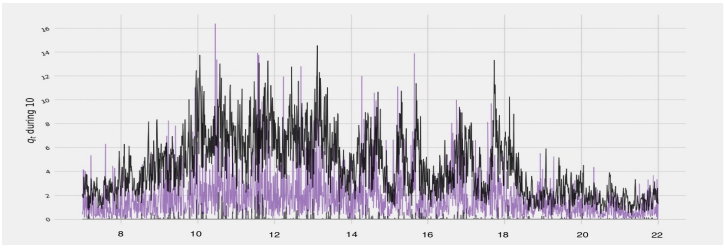
(a) CI of $\boldsymbol{q}^{\text{obs}}$ obtained with mean posterior estimators.

(b) CI of $\boldsymbol{q}^{\text{obs}}$ obtained by sampling posterior distribution of C_α and $\bar{\gamma}(t)$ from TI.

Figure 9: Comparison of the accuracy of the model to predict the instantaneous turbulent kinetic energy against the observation: the green plot is the instantaneous turbulent kinetic energy observed during February 15th, 2017, between 7 a.m and 8 p.m using the frequency of 1/30 s^{-1}. In Figure 9(a) (left), the black plots correspond to the confidence interval of $(\widehat{q}_{t_n}, n)$ in (5.1) using the posterior mean values $\widehat{C}_\alpha(d)$ and the time dependent mean $\bar{\gamma}(t) = \sum_{i=0}^{S} \mathbb{E}[\gamma(d,i)]\mathbb{1}_{[T_i,T_{i+1}]}(t)$, $\Delta t = 30$ s. In Figure 9(b) (right), the black plots correspond to the confidence interval of $(\widehat{q}_{t_n}, n)$ in (5.1) using the within-year posterior distribution of C_α and the time dependent mean $\bar{\gamma}(t)$ replaced by the turbulent intensity statistic through Equation (4.5), $\Delta t = 30$ s.



(a) $\boldsymbol{q}^{\text{obs}}$ for $d =$ `November` 1st and its predicted 95% CI



(b) $\boldsymbol{q}^{\text{obs}}$ for $d =$ `November` 10th and its predicted 95% CI

Figure 10: Prediction of the instantaneous turbulent kinetic energy: construction of 95% confidence intervals (in black) from $(\widehat{q}_{t_n}, n)$ in (5.1) using the within-year posterior distribution of C_α and the time dependent mean $\bar{\gamma}(t)$ replaced by the turbulent intensity statistic through Equation (4.5). Observations were taken during November 1st and November 10th between 7 a.m and 10 p.m (color plots).

precisely, we consider now the observed 10-minutes-averaged turbulent wind speed $\|u'^{\text{obs}}\| = \frac{1}{\xi}\sum_{t-\xi\leq s\leq t}\sqrt{q_s^{\text{obs}}}$ during all Wednesdays of 2017, with ξ =10 minutes, and we compute its empirical distribution. Similarly, we consider the modelled 10-minutes-averaged turbulent wind speed $\frac{1}{\xi}\sum_{t-\xi\leq s\leq t}\sqrt{\widehat{q}_s}$ using the numerical scheme (5.1) with C_α sampled from the posteriori distribution and γ_t given by the observed TI and Equation (5.2), and we compute its empirical distribution.

For a better comparison of the obtained distributions in Figure 11(a), we also adjust a model density to the empirical distributions. As mentioned in the introduction, within local forecasting, the Weibull probability distribution has been fitted to the average wind speed measurements. More precisely, the average wind speed over a 10 minutes-period is commonly considered as a Weibull random variable with shape parameter $k > 0$ and scale $\lambda > 0$ [7, 23]. The parameters k and λ can be estimated with different statistical methods, among which we can mention maximum likelihood methods, Anderson–Darling estimation of tails methods and with the Cramér–von Mises statistic [23].

The Weibull densities in Figure 11 have been fitted with simple estimators computed from the mode and median of the sample in order to illustrate density distributions. From this fit, we have obtained a Weibull distribution with shape parameter of $k_{\text{obs}} = 2.37$ and scale parameter of $\lambda_{\text{obs}} = 1.40$ for the observed turbulent wind speed, and shape parameter of $k_{\text{mod}} = 2.0$ and scale parameter of $\lambda_{\text{mod}} = 1.32$ for the approximation of the turbulent wind speed. Additionally, in Figure 11(b) we include a Q-Q plot comparing the approximated turbulent wind speed obtained from the CIR model against the observed one. Considering the comparison of the percentiles, we can conclude that the turbulent wind speed approximated from the proposed model fits very well to the Weibull distribution, showing a slight discrepancy only in the tails.

6. Conclusions

Starting from stochastic Lagrangian models for turbulent flows, we have modelled the turbulent kinetic energy filtered at a given point, introducing a production term denoted by γ. Remarkably, from the derived model, we recover the CIR-like process previously proposed in the literature, with our approach also providing a direct connection from the 3D+time physical model and parameters to the 0D+time model. The main feature of the production term γ (which has been proved analytically and numerically) is indeed to replicate the non-dissipative nature of the real wind data. We have proposed a complete calibration procedure (without external parameters) composed of a preliminary and a Bayesian inference stages. In the prior calibration (Step zero) we implemented the quadratic variation and maximum pseudo-likelihood estimators for γ and C_α, respectively. From these estimators we propose an a priori distribution for the vector parameter associated to the kinetic energy model (3.9) and estimate the observational error. Later, in the Bayesian step we quantify the uncertainty of the parameters by implementing Hamiltonian Monte Carlo methods within the construction of Markov chains converging to the stationary distribution of the model. This step provides a distribution for each parameter and allows variation in time for the parameter γ. The numerical study yields accurate results for both steps of the calibration, recovering the values of the physical parameter C_α, validating the model, quantifying the uncertainty and suggesting a model for the production term directly connected to the turbulence intensity statistic. Hence, if we have an initial guess on the turbulence intensity, (or if we separate the time interval into regimes with constant intensity), the model (3.9) with the closure (5.2) can be successfully used in the prediction of the turbulent kinetic energy.

In order to improve the replication of the fast oscillation in the observed process, a superlinear drift term can be considered in the model equation (3.9) with an exponential Euler scheme [16] for its discrete time version. For such drift with superlinear growth, the mean reverting term can act faster and the model can adapt better to the jumps present in the real data. This last point will be consider in a future work.

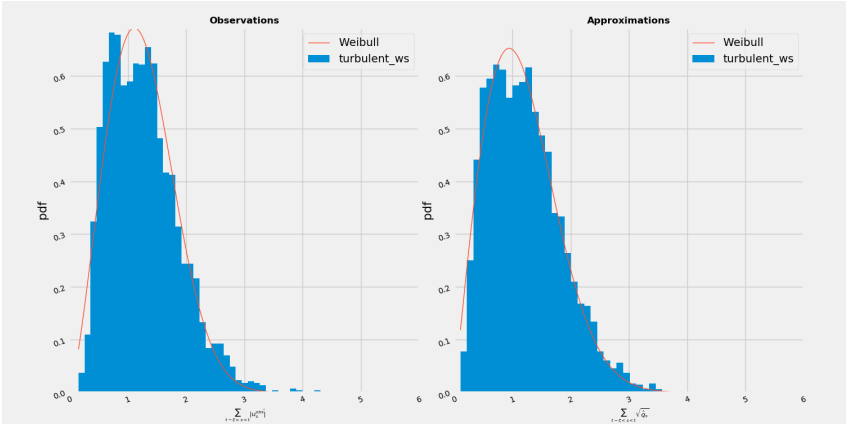
Acknowledgement

The authors would like to acknowledge SIRTA for providing high frequency data used in this study.
The second author acknowledges the support of the Russian Academic Excellence Project 5–100.
The third author acknowledges the support of ANID FONDECYT/POSTDOCTORADO N°321011.

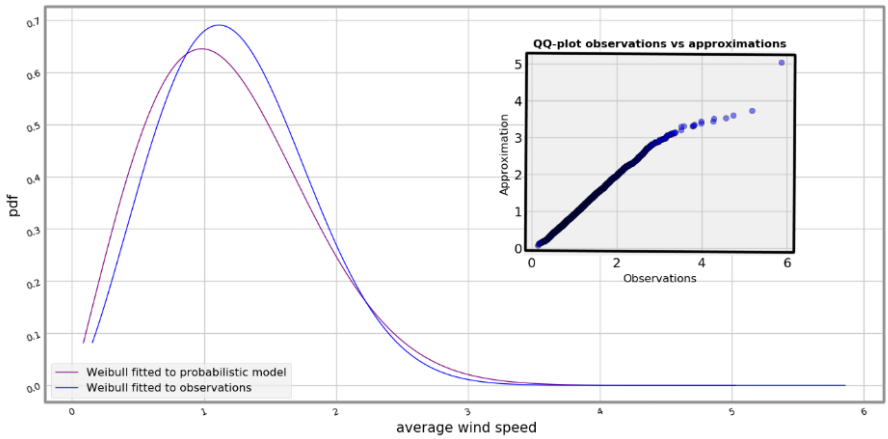
Appendix A. On the mean-field turbulent kinetic energy model

Appendix A.1. Proof of Proposition 3.2

By means of Lemma 3.4, the wellposedness of the mean-field model (3.4) can be easily deduced from the wellposedness of CIR processes with time-dependent coefficients. The latter can be proved with a time-change



(a) Histograms for the average turbulent wind speed.



(b) Comparison of the fitted densities and QQ-plot for the approximated turbulent wind speed against the observed turbulent wind speed.

Figure 11: Empirical density for the average turbulent wind speed computed with 10 minutes-period for all Wednesdays (2017) between 5 a.m and 8 p.m. Fig. 11(a) shows the empirical density corresponding to the observations $\frac{1}{\xi}\sum_{t-\xi\leq s\leq t}\sqrt{q_s^{\text{obs}}}$ (left) and the probabilistic model $\frac{1}{\xi}\sum_{t-\xi\leq s\leq t}\sqrt{\widehat{q}_s}$ (right). Figure 11(b) refines the comparison of the associated Weibull densities in Fig.11(a). Additionally, the graphical method of the quantile-quantile plot for observations against approximations is shown.

technique (see e.g. [37]). Indeed, let us consider the map $t \mapsto \widetilde{\mathrm{k}}(t)$ as the solution of the ordinary differential equation

$$\frac{d\widetilde{\mathrm{k}}(t)}{dt} = \gamma - \frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}^{3/2}(t), \qquad \widetilde{\mathrm{k}}_0 = \mathbb{E}[q_0], \tag{A.1}$$

and the extended CIR model:

$$dq_t = \gamma dt - C_R\frac{C_\alpha}{\sqrt{2}}q_t\widetilde{\mathrm{k}}(t)^{1/2}dt + 3C_0\frac{C_\alpha}{2\sqrt{2}}\widetilde{\mathrm{k}}(t)^{3/2}dt + \sqrt{\sqrt{2}C_0C_\alpha}\,\widetilde{\mathrm{k}}(t)^{3/4}\sqrt{q_t}dW_t, \quad q_0 > 0 \text{ given.} \tag{A.2}$$

Assuming $\gamma > 0$, by Lemma 3.4, we have that, for all $t \geq 0$:

$$\min\left\{q_0, \left(\tfrac{\sqrt{2}\gamma}{C_\alpha}\right)^{2/3}\right\} \leq \widetilde{\mathrm{k}}(t) \leq \max\left\{q_0, \left(\tfrac{\sqrt{2}\gamma}{C_\alpha}\right)^{2/3}\right\}, \tag{A.3}$$

and consequently, for all $t \geq 0$, we compute

$$\delta(t) := \frac{4\left(\gamma + 3C_0\frac{C_\alpha}{2\sqrt{2}}\widetilde{\mathrm{k}}(t)^{3/2}\right)}{\sqrt{2}C_0C_\alpha\,\widetilde{\mathrm{k}}(t)^{3/2}} = \frac{2\sqrt{2}\gamma}{C_0C_\alpha\widetilde{\mathrm{k}}(t)^{3/2}} + 3 \;> 2. \tag{A.4}$$

Then, from Theorem 2.5 in [37] there exists a unique strictly positive solution $(q_t; t \geq 0)$ to Equation (A.2). Now, considering this solution, we notice that $\mathbb{E}[q_t]$ is a non-negative solution to

$$\frac{d\mathbb{E}[q_t]}{dt} = F(t, \mathbb{E}[q_t]), \quad \text{with} \quad F(t,x) = \gamma - C_R\frac{C_\alpha}{\sqrt{2}}(\widetilde{\mathrm{k}}(t))^{1/2}x + 3C_0\frac{C_\alpha}{2\sqrt{2}}\widetilde{\mathrm{k}}(t)^{3/2}.$$

Estimate (A.3) ensures that $x \mapsto F(t,x)$ is uniformly Lipschitz continuous, and, since $\widetilde{\mathrm{k}}(t)$ in (A.1) satisfies the same equation as $\mathbb{E}[q_t]$, we have $\mathbb{E}[q_t] = \widetilde{\mathrm{k}}(t)$. This immediately guarantees that $(q_t; t \geq 0)$ is a solution to (3.4).

Appendix A.2. Proof of Lemma 3.4

For any $p \geq 1$, from Itô formula, applied on q_t^p with (3.4), we have

$$\mathbb{E}[q_t^p] = q_0^p + p\gamma\int_0^t \mathbb{E}[q_s^{p-1}]ds - pC_R\frac{C_\alpha}{\sqrt{2}}\int_0^t \mathbb{E}[q_s^p]\sqrt{\mathbb{E}[q_s]}ds + p(p+\tfrac{1}{2})C_0\frac{C_\alpha}{\sqrt{2}}\int_0^t \mathbb{E}^{3/2}[q_s]\mathbb{E}[q_s^{p-1}]ds.$$

By identifying $\widetilde{\mathrm{k}}_p(t) := \mathbb{E}[q_t^p]$, we deduce the following linear ordinary differential equation

$$\frac{d\widetilde{\mathrm{k}}_p(t)}{dt} = -pC_R\frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{1/2}(t)\,\widetilde{\mathrm{k}}_p(t) + p\big((p+\tfrac{1}{2})C_0\frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{3/2}(t) + \gamma\big)\widetilde{\mathrm{k}}_{p-1}(t) \tag{A.5}$$

with $\widetilde{\mathrm{k}}_1(t)$ the non-negative solution to

$$\frac{d\widetilde{\mathrm{k}}_1(t)}{dt} = (\gamma - \frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{3/2}(t))\mathbb{1}_{\{\widetilde{\mathrm{k}}_1(t)>0\}},$$

such that $\widetilde{\mathrm{k}}_1(t) = 0$ for all $t \geq \tau_1 := \inf\{t > 0 \,:\, \widetilde{\mathrm{k}}_1(t) = 0\}$.

In the case $\gamma = 0$. From Equation (A.5), we have $\frac{d\widetilde{\mathrm{k}}_1(t)}{dt} \leq 0$. Thus, the map $t \in [0, +\infty) \mapsto \widetilde{\mathrm{k}}_1(t)$ is non-increasing, leading immediately to the bound (3.7). We prove the bound (3.6), by applying a comparison principle on the ODE (A.5). Indeed, from Hölder inequality we know $\widetilde{\mathrm{k}}_1(t)\widetilde{\mathrm{k}}_{p-1}(t) \leq \widetilde{\mathrm{k}}_p(t)$, and then

$$\begin{aligned}-pC_R\frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{1/2}(t)\,\widetilde{\mathrm{k}}_p(t) + p(p+\tfrac{1}{2})C_0\frac{C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{3/2}(t)\widetilde{\mathrm{k}}_{p-1}(t) &\leq p\frac{C_0(p-1)C_\alpha}{\sqrt{2}}\widetilde{\mathrm{k}}_1^{3/2}(t)\widetilde{\mathrm{k}}_{p-1}(t)\\ &\leq -pC_0(p-1)\frac{d\widetilde{\mathrm{k}}_1}{dt}(t)\widetilde{\mathrm{k}}_{p-1}(t).\end{aligned}$$

Hence, for any $p \geq 1$, by means of a comparison principle we obtain

$$\widetilde{\mathrm{k}}_p(t) \leq q_0^p - pC_0(p-1)\int_0^t \frac{d\widetilde{\mathrm{k}}_1(s)}{dt}\widetilde{\mathrm{k}}_{p-1}(s)ds.$$

Iterating the previous inequality, and noticing that,

$$0 \leq -\int_0^t \frac{d\widetilde{\mathrm{k}}_1(s)}{dt} ds \leq q_0, \qquad \text{and} \qquad 0 \leq -\int_0^t \frac{d\widetilde{\mathrm{k}}_1(s)}{dt} \widetilde{\mathrm{k}}_1(s) ds \leq q_0^2/2,$$

we deduce (3.6) from

$$0 \leq \widetilde{\mathrm{k}}_p(t) \leq q_0^p \left(1 + \frac{1}{p(p-n)} \sum_{n=1}^{p-1} C_0^n \prod_{i=0}^{n} (p-i)^2 \right), \quad \text{for all } t \geq 0.$$

When $\gamma = 0$, Equation (A.5) can be explicitly solved for $p = 1$ with

$$\widetilde{\mathrm{k}}_1(t) = \mathbb{E}[q_t] = \left(q_0^{-1/2} + \frac{C_\alpha t}{2\sqrt{2}} \right)^{-2}. \tag{A.6}$$

Therefore $\tau_1 = +\infty$ and we conclude on (3.8) with the dissipation of the moment of the TKE at large times.

In the case $\gamma > 0$. Equation (A.5) can be written for $p = 1$ as:

$$C_{q_0} - \frac{C_\alpha}{\sqrt{2}} t = \frac{1}{\gamma'} \left\{ \frac{1}{3} \log \left(\frac{(\gamma' - \sqrt{\widetilde{\mathrm{k}}_1(t)})^2}{\widetilde{\mathrm{k}}_1(t) + \gamma' \sqrt{\widetilde{\mathrm{k}}_1(t)} + \gamma'^2} \right) + \frac{2}{\sqrt{3}} \arctan \left(\frac{2\sqrt{\widetilde{\mathrm{k}}_1(t)} + \gamma'}{\gamma' \sqrt{3}} \right) \right\}, \tag{A.7}$$

with $\gamma' = (\frac{\sqrt{2}\gamma}{C_\alpha})^{1/3}$ and C_{q_0} a constant depending on the initial condition. Then, we can analyse the limit behaviour of $\widetilde{\mathrm{k}}_1$ by means of the isocline method applied to the ODE (3.5) and the Equation (A.7). Indeed, from (A.7) we have $\widetilde{\mathrm{k}}_1(t) \neq (\frac{\sqrt{2}\gamma}{C_\alpha})^{2/3}$ for all $t \geq 0$, and then $\frac{d\widetilde{\mathrm{k}}_1(t)}{dt}$ is either strictly negative or strictly positive. Further, by computing the second derivative of $\widetilde{\mathrm{k}}_1(t)$:

$$\frac{d^2\widetilde{\mathrm{k}}_1}{dt^2}(t) = -\frac{3C_\alpha}{2\sqrt{2}} (\widetilde{\mathrm{k}}_1)^{1/2}(t) \frac{d\widetilde{\mathrm{k}}_1}{dt}(t),$$

we can check that $\widetilde{\mathrm{k}}_1$ does not change its curvature sign, and

$$\widetilde{\mathrm{k}}_1(t) > \left(\frac{\sqrt{2}\gamma}{C_\alpha} \right)^{2/3} \text{ implies } \quad \frac{d\widetilde{\mathrm{k}}_1}{dt}(t) < 0 \text{ and } \frac{d^2\widetilde{\mathrm{k}}_1}{dt^2}(t) > 0.$$

Thus, in the set $\{t : \widetilde{\mathrm{k}}_1(t) > (\frac{\sqrt{2}\gamma}{C_\alpha})^{2/3}\}$, the map $t \mapsto \widetilde{\mathrm{k}}_1(t)$ decreases with time and is convex. Likewise, when $\widetilde{\mathrm{k}}_1(t) < (\frac{\sqrt{2}\gamma}{C_\alpha})^{2/3}$, the map $t \mapsto \widetilde{\mathrm{k}}_1(t)$ increases and is concave. We conclude on the boundedness of $\widetilde{\mathrm{k}}(t)$ and on its behaviour at large times:

$$\lim_{t \to +\infty} \widetilde{\mathrm{k}}_1(t) = \left(\frac{\sqrt{2}\gamma}{C_\alpha} \right)^{2/3}.$$

Similarly to the case $\gamma = 0$, for any $p > 1$ we can bound the right hand-side term in Equation (A.5) in order to show that the pth-moment is uniformly bounded. This time, we use the non-null lower bound of $\widetilde{\mathrm{k}}_1$: $\min\{q_0, (\frac{\sqrt{2}\gamma}{C_\alpha})^{2/3}\}$, obtaining:

$$\begin{aligned} -\underline{\xi}_{1,p} \widetilde{\mathrm{k}}_p(t) + \underline{\xi}_{2,p} \widetilde{\mathrm{k}}_{p-1}(t) &\leq -pC_R \frac{C_\alpha}{\sqrt{2}} \widetilde{\mathrm{k}}_1^{1/2}(t) \widetilde{\mathrm{k}}_p(t) + p\big((p + \tfrac{1}{2})C_0 \frac{C_\alpha}{\sqrt{2}} \widetilde{\mathrm{k}}_1^{3/2}(t) + \gamma\big) \widetilde{\mathrm{k}}_{p-1}(t) \\ &\leq -\overline{\xi}_{1,p} \widetilde{\mathrm{k}}_p(t) + \overline{\xi}_{2,p} \widetilde{\mathrm{k}}_{p-1}(t), \end{aligned}$$

for some positive constants $\overline{\xi}_{1,p}, \overline{\xi}_{2,p}, \underline{\xi}_{1,p}$ and $\underline{\xi}_{2,p}$ depending on p, C_α, C_0, γ and C_R. Then, from an induction argument on p and a comparison principle we deduce that

$$\begin{aligned} \frac{\underline{\xi}_{2,p}}{\underline{\xi}_{1,p}} \inf_{s \geq 0} \widetilde{\mathrm{k}}_{p-1}(s) - \left(\frac{\underline{\xi}_{2,p}}{\overline{\xi}_{1,p}} \inf_{s \geq 0} \widetilde{\mathrm{k}}_{p-1}(s) - q_0^p \right) \exp\left\{ -\underline{\xi}_{1,p}\, t \right\} \\ \leq \widetilde{\mathrm{k}}_p(t) \leq \frac{\overline{\xi}_{2,p}}{\overline{\xi}_{1,p}} \sup_{s \geq 0} \widetilde{\mathrm{k}}_{p-1}(s) - \left(\frac{\overline{\xi}_{2,p}}{\overline{\xi}_{1,p}} \sup_{s \geq 0} \widetilde{\mathrm{k}}_{p-1}(s) - q_0^p \right) \exp\left\{ -\overline{\xi}_{1,p}\, t \right\}, \end{aligned}$$

for all $t \geq 0$, i.e. $\widetilde{\mathrm{k}}_p$ is bounded uniformly in time, which ends the proof.

References

[1] J.P. Arenas-López and M. Badaoui. The Ornstein-Uhlenbeck process for estimating wind power under a memoryless transformation. *Energy*, 213:118842, 2020. 2

[2] J.P. Arenas-López and M. Badaoui. Stochastic modelling of wind speeds based on turbulence intensity. *Renewable Energy*, 115:10–22, 2020. 2, 7

[3] J. Badosa, E. Gobet, M. Grangereau, and D. Kim. Day-ahead probabilistic forecast of solar irradiance: a stochastic differential equation approach. In *Forecasting and Risk Management for Renewable Energy*, pages 73–93. Springer, 2017. 7

[4] C. Baehr. Nonlinear filtering for observations on a random vector field along a random path. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44:921–945, 2010. 3, 7

[5] C. Baehr, C. Beigbeder, F. Couvreux, A. Dabas, and B. Piguet. Retrieval of the turbulent and backscattering properties using a nonlinear filtering technique applied to Doppler LIDAR observation. In *16th Int. Symp. for the Advancement of Boundary-Layer Remote Sensing (ISARS)*, 2012. 7

[6] A. Bensoussan and A. Brouste. Cox–Ingersoll–Ross model for wind speed modeling and forecasting. *Wind Energy*, 19(7):1355–1365, 2016. 2, 7, 10

[7] A. Bensoussan, P. Bertrand, and A. Brouste. A generalized linear model approach to seasonal aspects of wind speed modeling. *Journal of Applied Statistics*, 41(8):1694–1707, 2014. 25

[8] A. Berkaoui, M. Bossy, and A. Diop. Euler scheme for SDEs with non-Lipschitz diffusion coefficient: strong convergence. *ESAIM: Probability and Statistics*, 12:1–11, 2008. 13

[9] F. Bernardin, M. Bossy, C. Chauvin, P. Drobinski, A. Rousseau, and T. Salameh. Stochastic downscaling methods: Application to wind refinement. *Stochastic Environmental Research and Risk Assessment*, 23(6):851–859, 2009. 3, 7

[10] F. Bernardin, M. Bossy, C. Chauvin, J.-F. Jabir, and A. Rousseau. Stochastic Lagrangian method for downscaling problems in computational fluid dynamics. *ESAIM: Mathematical Modelling and Numerical Analysis, EDP Sciences. Special Issue on Probabilistic methods and their applications*, 44(5):885–920, 2010. 3, 5

[11] M. Bossy. Some stochastic particle methods for nonlinear parabolic PDEs. volume 15, pages 18–57, 2005. 9

[12] M. Bossy and A. Diop. Weak convergence analysis of the symmetrized Euler scheme for one dimensional SDEs with diffusion coefficient $|x|^\alpha,\ \alpha \in [\frac{1}{2}, 1)$. *Inria Research Report RR-5396. Available at arxiv:1508.04573.R*, 2004. 13

[13] M. Bossy and H. Olivero. Strong convergence of the symmetrized Milstein scheme for some CEV-like SDEs. *Bernouilli*, 24(3):1995–2042, 2018. 13

[14] M. Bossy, J. Espina, J. Morice, C. Paris, and A. Rousseau. Modeling the wind circulation around mills with a Lagrangian stochastic approach. *SMAI Journal of Computational Mathematics*, 2:177–214, 2016. 3, 5

[15] M. Bossy, A. Dupré, P. Drobinski, L. Violeau, and C. Briard. Stochastic Lagrangian approach for wind farm simulation. In *Forecasting and Risk Management for Renewable Energy*, pages 45–71. Springer International Publishing, 2018. ISBN 978-3-319-99052-1. 3, 5

[16] M. Bossy, J-F. Jabir, and K. Martínez. On the weak convergence rate of an exponential Euler scheme for SDEs governed by coefficients with superlinear growth. *Bernoulli*, 27(1):312–347, 2021. ISSN 1350-7265. doi: 10.3150/20-BEJ1241. 25

[17] J.A. Carta, P. Ramírez, and S. Velázquez. A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands. *Renewable and sustainable energy reviews*, 13(5):933–955, 2009. 7

[18] G. Casella and R. Berger. *Statistical Inference*. Second Edition. Duxbury Advanced Series, 2002. 13

[19] W.Y. Chang. A literature review of wind forecasting methods. *Journal of Power and Energy Engineering*, 2(04):161–168, 2014. 2

[20] J. Cox, J. Ingersoll, and S. Ross. A theory of the term structure of interest rates. *Econometrica*, 53(2):385–407, 1985. 10

[21] S. Dereich, A. Neuenkirch, and L. Szpruch. An Euler-type method for the strong approximation of the Cox–Ingersoll–Ross process. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 468(2140):1105–1115, 2012. 10

[22] T. Dreeben and S.B. Pope. Probability density function/Monte Carlo simulation of near-wall turbulent flows. *Journal of Fluid Mechanics*, 357:141–166, 1998. 6

[23] P. Drobinski, C. Coulais, and B. Jourdier. Surface wind-speed statistics modelling: alternatives to the Weibull distribution and performance evaluation. *Boundary-Layer Meteorology*, 157(1):97–123, 2015. 25

[24] P.A. Durbin. A Reynolds stress model for near-wall turbulence. *Journal of Fluid Mechanics*, 249:465–498, 1993. 6

[25] P.A. Durbin and B.A. Reif-Pettersson. The elliptic relaxation method. *Closure Strategies for Turbulent and Transitional Flows*, pages 127–152, 2002. 6

[26] P.A. Durbin and C-G. Speziale. Realizability of second-moment closure via stochastic analysis. *Journal of Fluid Mechanics*, 280:395–407, 1994. 6, 8

[27] W.N. Edeling, P. Cinnella, R.P. Dwight, and H. Bijl. Bayesian estimates of parameter variability in the $k-\varepsilon$ turbulence model. *Journal of Computational Physics*, 258(C):73–94, 2014. doi: https://doi.org/10.1016/j.jcp.2013.10.027. 4, 13, 15

[28] A. Gelman and M. Hoffman. The No–U–Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15:1593–1623, 2014. 16

[29] T. Göçmen and G. Giebel. Estimation of turbulence intensity using rotor effective wind speed in Lillgrund and Horns Rev-I offshore wind farms. *Renewable energy*, 99:524–532, 2016. 11

[30] M. Haeffelin, L. Barthès, O. Bock, C. Boitel, S. Bony, D. Bouniol, H. Chepfer, M. Chiriaco, J. Cuesta, J. Delanoë, P. Drobinski, J.-L. Dufresne, C. Flamant, M. Grall, A. Hodzic, F. Hourdin, F. Lapouge, Y. Lemaître, A. Mathieu, Y. Morille, C. Naud, V. Noël, W. O'Hirok, J. Pelon, C. Pietras, A. Protat, B. Romand, G. Scialom, and R. Vautard. SIRTA, a ground-based atmospheric observatory for cloud and aerosol research. *Annales Geophysicae*, 23(2):253–275, 2005. doi: 10.5194/angeo-23-253-2005. 10

[31] S. Hanifi, X. Liu, Z. Lin, and S. Lotfian. A critical review of wind power forecasting methods—past, present and future. *Energies*, 13(15): 3764, 2020. 2

[32] A. Honrubia, A. Vigueras-Rodríguez, and E. Gómez-Lázaro. *The Influence of Turbulence and Vertical Wind Profile in Wind Turbine Power Curve*, volume IV of *Progress in Turbulence and Wind Energy, SPPHY 141*. Springer, 2000. 11

[33] N. Ikeda and S. Watanabe. *Stochastic differential equations and diffusion processes*. North-Holland Publishing Company, 1981. 9

[34] P-L. Jabin and Z. Wang. *Active Particles, Volume 1. Advances in Theory, Models and Applications*, chapter Mean field limit for stochastic particle systems, pages 379–402. Birkhauser-Springer, Boston, 2017. 9

[35] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Probability Theory and Stochastic Processes, 2nd edition. Springer-Verlag, Berlin, 1991. 10

[36] M.K. Lauren, M. Menabde, and G.L. Austin. Analysis and simulation of surface-layer winds using multiplicative cascade models with self-similar probability densities. *Boundary-layer meteorology*, 100(2):263–286, 2001. 2

[37] Y. Maghsoodi. Solution of the extended CIR term structure and bond option valuation. *Mathematical Finance*, 6(1):89–109, 1996. 27
[38] R. Manceau and K. Hanjalić. Elliptic blending model: A new near-wall Reynolds-stress turbulence closure. *Physics of Fluids*, 14(2): 744–754, 2002. 6
[39] W. C. Miao. *Quadratic variation estimators for diffusion models in finance*. PhD thesis, 2004. 14
[40] J-P. Minier. Statistical descriptions of polydisperse turbulent two-phase flows. *Physics Reports*, 665:1 – 122, 2016. ISSN 0370-1573. 3
[41] J-P. Minier and E. Peirano. The pdf approach to turbulent polydispersed two-phase flows. *Physics Reports: A Review Section of Physics Letters*, 352:1–214, 2001. 3
[42] A. Murata, H. Ohtake, and T. Oozeki. Modeling of uncertainty of solar irradiance forecasts on numerical weather predictions with the estimation of multiple confidence intervals. *Renewable energy*, 117:193–201, 2018. 7
[43] R.M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2, 2011. 16
[44] J.T. Oden, R. Moser, and O. Ghattas. Computer predictions with quantified uncertainty, Part II. *SIAM News*, 43(10):1–4, 2010. 4
[45] S.B. Pope. Pdf methods for turbulent reactive flows. *Progress in Energy and Combustion Science*, 11(2):119 – 192, 1985. ISSN 0360-1285. 3, 5
[46] S.B. Pope. On the relationship between stochastic Lagrangian models of turbulence and second-moment closures. *AIP: Physics of Fluids*, 973(6), 1994. 5, 6
[47] S.B. Pope. Lagrangian pdf methods for turbulent flows. *Annu. Rev. Fluid Mech.*, 26:23–63, 1994. 3, 6
[48] S.B. Pope. *Turbulent flows*. Cambridge University Press, 2000. 2, 5, 6, 7
[49] P. Protter. *Stochastic Integration and Differential Equations*. Stochastic Modeling and Applied Probability. Springer–Verlag, Berlin, second edition edition, 2004. 8
[50] L. Rottner, C. Baehr, A. Dabas, and L. Hammoud. Stochastic method for turbulence estimation from Doppler lidar measurements. *Journal of Applied Remote Sensing*, 11(4):046001, 2017. 7
[51] John Salvatier, Thomas V Wiecki, and Christopher Fonnesbeck. Probabilistic programming in Python using PyMC3. *PeerJ Computer Science*, 2:e55, 2016. 17
[52] S. Soman, H. Zareipour, O. Malik, and P. Mandal. A review of wind power and wind speed forecasting methods with different time horizons. In *North American Power Symposium 2010*, pages 1–8. IEEE, 2010. 2
[53] A. Stohl. Computation, accuracy and applications of trajectories—A review and bibliography. *Atmospheric Environment*, 32(6):947 – 966, 1998. ISSN 1352-2310. 3
[54] F. Suzat, C. Baehr, and A. Dabas. A fast atmospheric turbulent parameters estimation using particle filtering. Application to LIDAR observations. In *J. Phys.: Conf. Ser.*, volume 318, page 072019, 2011. 3, 7
[55] C.Y. Tang and S.X. Chen. Parameter estimation and bias correction for diffusion processes. *Journal of Econometrics*, 149(1):65 – 81, 2009. 14
[56] M. Wacławczyk, J. Pozorski, and J-P. Minier. Probability density function computation of turbulent flows with a new near-wall model. *Physics of Fluids*, 16(5):1410–1422, 2004. doi: 10.1063/1.1683189. 6