Learning-based Emulation of Sea Surface Wind Fields from Numerical Model Outputs and SAR Data

Liyun He-Guelton, Ronan Fablet, Bertrand Chapron, Jean Tournadre

Learning-based Emulation of Sea Surface Wind Fields from Numerical Model Outputs and SAR Data

Liyun He, Ronan Fablet, Bertrand Chapron, and Jean Tournadre

***Abstract*—The availability of sea surface wind conditions with a high-resolution space-time sampling is a critical issue for a wide range of applications. Currently, no observation systems nor model forecasts provide relevant information with a high sampling rate both in space and time. Synthetic Aperture Radar (SAR) satellite systems deliver high-resolution sea surface fields, with a spatial resolution below 0.01°, but they are also characterized by a large revisit time up 7-to-10 days for temperate zones. Meanwhile, operational model predictions typically involve a high temporal resolution (*e.g.* every $6\,\mathrm{h}$), but also a low spatial resolution (0.5°). With a view to leveraging both data sources, we investigate statistical downscaling schemes. In this study, a new model based on a machine learning method, namely Support Vector Regression (SVR), is built to reconstruct high-resolution sea surface wind fields from low-resolution operational model forecasts. The considered case study off Norway demonstrates the relevance of the proposed SVR model. It outperforms state-of-the-art approaches (namely, linear, analog and Empirical Orthogonal Function (EOF) downscaling models) in terms of mean square error. It also realistically reproduces complex space-time variabilities of the observed SAR wind fields. We further discuss the SVR model as a generalization of the popular linear and analog models.**

***Index Terms*—Machine learning, Downscaling, Coastal wind, High-resolution, Support vector regression (SVR)**

I. Introduction

THE derivation of local fine-scale information from coarse-resolution conditions provided by general circulation models using statistical models is generally referred to as statistical downscaling [1], [2]. Statistical downscaling provides a mean to solve for the scale mismatch between numerical model predictions and satellite observations. Whereas numerical models typically involve time resolutions up to a few hours, they provide rather coarse predictions in space. For instance, ECMWF analyses sea surface wind fields are associated with a 0.5° (or 0.25°) and 6-hour (or 3-hour) space-time sampling. By contrast, satellite observations can provide High Resolution (HR) sea surface geophysical fields, up to a resolution of 0.01° for SAR sea surface wind fields [3], [4]. However, satellite SAR systems involve a highly irregular sampling of the ocean surface and, for a given region, SAR wind fields may be delivered with a low temporal resolution, typically every 7-to-10 days for temperate zones. It makes them not adequate for direct use in operational weather forecasting or for assimilation into numerical forecast models. Thus we seek to explore a statistical downscaling strategy to benefit both from the high spatial resolution of the SAR-derived fields, and the regular time resolution of the numerical model outputs.

Due to its relative simplicity and low computational costs, statistical downscaling is particularly appealing to leverage these two sources of data [1], [5]. It is particularly useful for a heterogeneous environment with a complex geography and topography, involving for instance islands and mountains. For such conditions, the physical processes are difficult to model directly as involved in dynamical downscaling [6]. By contrast, statistical downscaling is regarded as a generic and tractable learning-based strategy to calibrate a downscaling model from training datasets, as illustrated for large-scale climate information in [7].

In this study, we address the synergy between numerical model predictions and satellite observations to deliver HR space-time predictions of sea surface geophysical fields. Formally, we investigate statistical downscaling techniques, which are stated as regression issues. Statistical downscaling has been initially developed for local precipitation prediction [8], [2]. More recently, applications to sea surface parameters, such as sea salinity, sea level and sea temperature, have also been dealt with [9], [10]. From a methodological point of view, linear and non-linear regression techniques have been evaluated. For instance, [9], [10] compare linear regression method with neural networks, a non-linear regression model, for the reconstruction of Sea Surface Temperature (SST) anomalies from Low Resolution (LR) sea level pressure and SST conditions. Analog regression has also received a great attention. For precipitation processes, Zorita *et al.* [2] show that the analog method generally performs as well as more complex methods for rainfall prediction. From a machine learning point of view, the last decades have seen the emergence of novel learning-based regression techniques, among which Support Vector Regression (SVR) [11], Random Forest [12] and Neural Networks [13] are the most popular and powerful. Here, we focus on SVR, which can be regarded as a generalization of both analog and linear regression schemes [14].

Along with the regression models, the selection of the regression variables is a key issue. In most statistical downscaling applications, global representations of the LR and HR fields are issued from their projection onto Empirical Orthogonal Functions (EOFs) learned from the training data [15], [2], [9], [10], [16], [17]. The EOF scheme extracts an orthonormal transformation to represent the considered geophysical fields, *e.g.* all the wind vectors in the study area, as a low-dimensional feature vector. This dimension reduction

greatly simplifies the learning step of the downscaling issue. However, it may lead to information loss, which may be critical to actually account for local effects. Local regression models then appear as appealing solutions and we investigate their development and relevance in this study.

This paper is organized as follows. Section II presents the considered case study, including the associated data and key geophysical patterns of the study area. The proposed learning-based approach is described in Section III, along with the comparison to the state-of-the-art approaches. Section IV reports and discusses numerical experiments, in terms of both emulation performance and geophysical patterns. We further discuss key features and future work in section V.

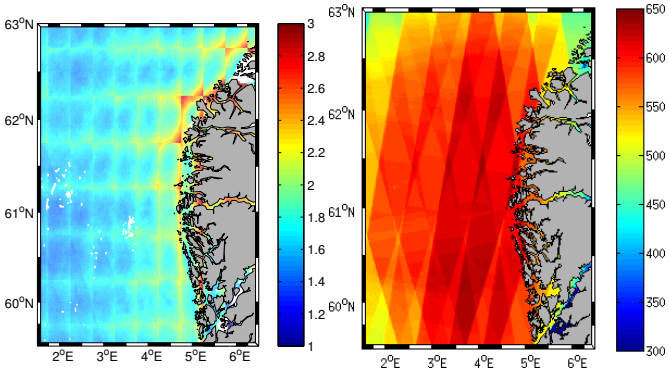
II. Data and Study Area

This study addresses the reconstruction of HR sea surface wind fields at 10 m height. HR wind fields are of great interest for a wide range of applications [3]. They clearly contribute to the understanding of wind field dynamics, especially in coastal areas, and the improvement of numerical prediction models. Among others, the availability of HR wind data will help the assessment of energy production, risks relevant to marine engineering, environment pollution, security, *etc*.

HR SAR data, issued from the ENVironmental SATellite (ENVISAT) and processed by the Collecte Localisation Satellites (CLS) centre, reach a spatial resolution of $0.01°$, corresponding to about $1.10\,\text{km} \times 0.55\,\text{km}$ at the latitude of about $60°$ [3]. SAR is known for its improved spatial resolution compared to current scatterometers (typically $0.25°$). It makes it particularly useful in coastal areas, which involves complex fine-scale dynamics. SAR wind speeds are routinely estimated using an empirical Geophysical Model Function (GMF), such as C-band model 4 (CMOD4, [18]), C-band model 5 (CMOD5, [19]). They relate wind vectors to the measured normalized radar cross section (NRCS). Along with the computation of the wind speed, the determination of the wind direction is a key issue. Zhang *et al.* [20] give an overview of existing methods to get this information. The most common approach consists in using wind direction given by an ancillary source of information from numerical weather prediction (NWP) models [21] or other remote sensing instrument such as scatterometer [22]. CLS SAR wind products use the European Center for Medium-range Weather Forecast (ECMWF) wind directions to initialize the wind retrievals from SAR imagery. The SAR-derived wind vectors are determined by a Bayesian estimator from normalized radar cross section (NRCS) measurements [4]. The mean error in SAR wind speed and direction is typically less than $2.0\,\text{m}\,\text{s}^{-1}$ and $25°$ respectively [4].

We exploit as LR data the analysis data delivered by the ECMWF, with a spatial resolution of $0.5°$. These data are available every 6 h (0 h, 6 h, 12 h, 18 h UTC). According to the scale categorisation of [23], ECMWF data relate to the meso-scale (at time scale from 1 h to a few days and for spatial scales from a few kilometers to a few thousand kilometers). By contrast, SAR data can reveal micro-scale patterns, associated with more local effects, such as gravity waves, barrier jets, turbulence, land-sea breezes, *etc*.

The study area is an area off the southwest coast sea of Bergen. It involves particularly complex sea wind conditions due to the presence of many mountains, islands and fjords [24]. The induced variability of LR-HR relationships [25] makes it an interesting study area for the targeted downscaling of HR wind fields. In this respect, we report a statistical analysis of the LR-HR relationships in the study area. Whereas ECMWF and SAR data are very alike in the offshore area (Figure 1a). Increasing differences are observed when getting closer to the coast. Such discrepancies in coastal areas are due to the coarser resolution of ECMWF data that prevents the model from accounting for the local changes in topography and surface land-sea roughness, as well as to the fact that small-scale features of non-homogeneous winds, such as sharp gradients due to atmospheric fronts, coastal jets and wind shadows, cannot be represented with coarse numerical simulations [26], [25]. Figure 1a further stresses that coarse-scale-to-fine-scale relationships vary from one point to another. As a consequence, learning point-specific regression model, rather than a global regression model, appears as a relevant choice.



(a) (b)

Fig. 1: **Considered case-study area:** root-mean-square deviation in $\text{m}\,\text{s}^{-1}$ between ECMWF and SAR data (a) and **number of available SAR-ECMWF pairs at each grid point in the given study area** (b). A SAR-ECMWF pair is formed by SAR wind field and the temporally closest ECMWF field. The SAR-ECMWF time difference remains lower than 3 h. In each grid point, the root-mean-square deviation is calculated with all available SAR wind fields acquired from 2005 to 2010 and the temporally closest ECMWF fields. The overpass time of ENVISAT in Bergen coast sea is around 21h30 UTC for ascending passes and around 10h UTC for descending passes.

As a complementary illustration, we report the wind distribution at a fjord point (N62.23°, E5.90°) (Figure 2). ECMWF and SAR data clearly depict very different wind distributions. In relation to the main orientation of the fjord, the SAR data involves a clear south, south-west and north-east dominant wind direction, whereas no such dominant pattern is observed from ECMWF data (wind roses in Figure 2). These empirical observations strongly motivate the exploitation of non-linear models to account for the non-linear relationship between coarse-scale ECMWF wind data and fine-scale wind data.

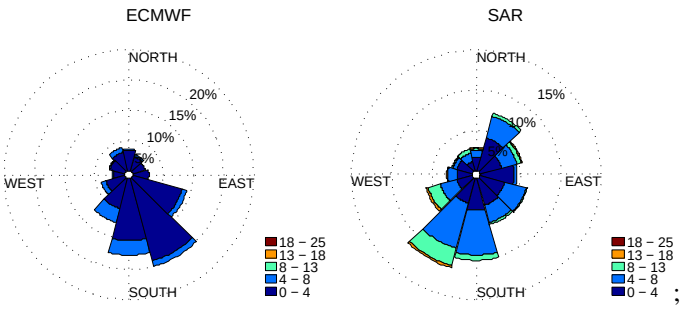


Fig. 2: **Wind roses at point (N**$62.23°$**, E**$5.90°$**)**: this fjord point corresponds to location 4 in Figure 5. It accounts for a inner fjord conditions with surrounding mountains. We compare ECMWF (left) and SAR (right) wind statistics. Overall, for this grid point, 520 SAR-ECMWF data pairs are available.

We have collected a dataset of 860 pairs of ECMWF and SAR wind field data in the area N$59°50'$ – N$63°0'$ and E$1°50'$ – E$6°50'$. The ENVISAT SAR data were acquired from 2005 to 2010. Each SAR data is co-located with the temporally closest ECMWF field. Hence, the time difference remains lower than 3 h. Pairs of ECMWF and SAR data that are very different are withdrawn from the analysis. The similarity measure between two ECMWF and SAR wind fields is evaluated as the mean square difference between the two fields in offshore zone. Empirically, a maximum relative difference of 0.4 was proven relevant to balance between the consistency of each ECMWF-SAR pair and the representativeness of the ECMWF-SAR dataset. Overall, our dataset comprises 758 ECMWF-SAR pairs.

For a given region, a single SAR image can not always cover the whole area whereas the ECMWF data are available everywhere. At each HR grid point, we use just the SAR-ECMWF pairs when SAR wind vector is available. Figure 1b illustrates the distribution of the number of the available SAR-ECMWF pairs at each grid point. [27] conclude that the database has a good representation of different wind conditions when the number of data is above 600. However, this is not always the case for border and fjord grid points. Because of this, the statistical errors of the downscaling method at these grid points may be relatively higher than at other locations.

III. Learning-based downscaling

In this section, we introduce the proposed learning-based approach. The reconstruction of a HR field from a LR model prediction is stated as a regression problem. Let us denote by y the HR field and by x LR field. The problem is modeled as

$$y = f(x) \tag{1}$$

where f is the regression function. Fields x and y are two-dimensional vector fields parameterized according to the zonal and meridional wind components.

Given a set of training data $\{(x_i, y_i)\}$, the learning task resorts to identifying the optimal regression function $f^* \in \mathcal{F}$. Within the solution space $\mathcal{F}$, the learning step comes to the minimization of some cost function L, for instance the mean square error. This general regression framework involves two key-elements: the definition of regression model f and the definition of regression variables issued from field x. The relevance and performance of the model depends on both elements. In this section, we first review the linear and analog regression models. We then introduce the SVR model, a state-of-the-art machine learning framework which can be viewed as a generalization of the two other models. This section also addresses the definition of regression variables as well as learning and implementation issues.

A. Linear regression

Due their simplicity, linear models are widely used in the downscaling context [28], [17], [29]. A linear regression parameterizes as:

$$y = A \cdot x \tag{2}$$

Here variable x is assumed to include a one-valued feature to account for a constant term in the linear regression. Given a training dataset $\{(x_i, y_i)\}$, the least-square estimation of matrix A resorts to:

$$A = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{3}$$

where the $n \times n$ matrix $\mathbf{X}^T\mathbf{X}$ consists of inner products between pairs of samples x_i, x_j and $\mathbf{y}$ is the vector of corresponding n-output. It might be noted that linear regression (2) may then be rewritten as a linear combination over training samples:

$$y = \sum_i \lambda_i \langle x, x_i \rangle y_i \tag{4}$$

where coefficients $\{\lambda_i\}$ directly relate to estimated matrix A [14].

B. Analog regression

The analog regression is another popular regression model for downscaling applications. It has been essentially applied in the field of weather forecasting [30], [2]. Given a new LR sample x, the key idea is to retrieve the data of the training data the closest to x to predict the HR field from the previously observed HR fields. It resorts to the following parameterization of the regression function:

$$f(x) = \sum_{s=1}^{n} w_s \, \mathrm{g}(x, x_s) y_s \tag{5}$$

where $\mathrm{g}(x, x_s)$ is a similarity measure between the input variable vector x and the s^{th} sample x_s. Weights $\{w_s\}$ are set *a priori* by users [2]. They are typically computed as a constant normalization factor:

$$w_s = \frac{1}{\sum_{s=1}^{n} \mathrm{g}(x, x_s)} \tag{6}$$

Nearest-neighbor and K-Nearest Neighbors regression models are specific parameterization of the analog regression with a binary parameterization of the similarity measure [2]. One may also considered other types of similarity measure such as Gaussian function (also referred to as radial basis function) [14]. It may be noted that when considering similarity measure associated with mean square difference, the selection

of EOF components as input variables x can be regarded as a mean to reduce the computational complexity of the evaluation of the distance in the original higher-dimensional space spanned by field x.

C. Support Vector Regression

In addition to the analog and Multiple Linear Regression (MLR) methods used as benchmarked downscaling schemes [2], [28], [17], we investigate in this study an optimal non-linear kernel-based regression model, namely SVR. The Support Vector Machine (SVM) and SVR (SVM for regression) have become particularly popular for nonlinear classification and regression [31], [32].

The SVR can be regarded as a linear regression model in a space defined by a non-linear mapping function Φ [33]:

$$f(x,\omega) = \omega^t \Phi(x) + b \tag{7}$$

where ω is a weight vector and b is the bias. The key idea of this formulation is to use a non-linear mapping Φ to project the data to another space where there exists a linear solution to the problem. In ε-SVR regression, the calibration of the regression model, *i.e.,* the calibration of weights $\{\omega_i\}$ and bias b resorts to minimising:

$$\frac{1}{2}\|\omega\|^2 + \mathrm{C}\sum_{i=1}^{n}(\xi_i + \xi_i^*) \tag{8}$$

under the following constraints:

$$\begin{cases} y_i - \langle \omega, \Phi(x_i)\rangle - b \leq \varepsilon + \xi_i \\ \langle \omega, \Phi(x_i)\rangle + b - y_i \leq \varepsilon + \xi_i^* \qquad \forall i \in (1,n) \\ \varepsilon, \xi_i, \xi_i^* \geq 0 \end{cases} \tag{9}$$

where ε is an accuracy parameter. Errors between observations $\{y_i\}$ and predictions $\langle \omega, \Phi(x_i)\rangle + b$ smaller than ε are ignored. Slack variables ξ_i and ξ_i^*, which are error measurements above and below the ε-insensitivity zone respectively, correspond to the soft margin. Regularization parameter C determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated [31].

The key feature of the SVR model is that there is no need for an explicit knowledge of mapping function Φ. The regression model can be rewritten according to the kernel function K, defined by $K(x_i, x) = \langle \Phi(x_i), \Phi(x)\rangle$, as [34]:

$$f(x,\omega) = \sum_i \alpha_i K(x_i, x) + b \tag{10}$$

where α_i sets the relative weight of each training data in the regression model. Hence, given a kernel, the training of the SVR model resorts to the inference of the weight vector according to margin-based criterion (Eq.8). Another particularly important property of the SVR model is the sparsity of the weights α_i. It can be shown and verified experimentally that only few weights are non-zero, meaning that the sum in Eq.10 reduces to a sum over only a few training samples.

The SVR model involves hyperparameters C or ε as well as the choice and parameterization of the kernel function. Here, we will consider a radial basis function with scale parameter γ. The calibration of these hyperparameters typically involves an exhaustive search over a grid of hyperparameter values using cross-validation statistics in terms of mean square error [35].

As summarized in Table I, the SVR setting (Eq.10) provides a generalization of the linear regression in Eq.4, as non-linear relationships may be accounted for. It also generalises the analog regression. Whereas the analog regression involves an empirical parameterization of the weighted regression function as in Eq.6, the SVR determines optimal weighing factors according to a margin-based criterion (Eq.8).

TABLE I: Synthesis of the different regression models used in this study.

Regression model	Support data choices $\{x_s\}$	Model calibration	Kernel function K
Analog	*All training samples*	*"Expert-base" setting*	*Any similarity function*
MLR	*All training samples or Centroids*	*Least-Square criterion*	*linear kernel* $\langle x, x_s\rangle$
SVR	*Support Vectors*	*Margin-based criterion*	*Mercer kernel [34]*

D. Definition of regression variables

For the targeted application, the input variables are issued from the LR ECMWF sea surface winds. Different approaches may be considered. Following previous work [15], [10], [16], [17], one may exploit a global low-dimensional EOF-based representation of the ECMWF wind field. Point-specific regression variables may also be investigated, given the spatial variability highlighted by the analysis performed in Section II. In addition to a global EOF-based representation, we evaluate two other schemes:

Local information It consists in exploiting the LR wind information within a local neighborhood around the HR grid point. It requires determining the optimal window size W_p (yellow box in Figure 3).

Entropy-based information One may also aim at locally selecting the best regression variables for a given HR point. As an exhaustive search over all possible variable sets is not feasible in practice, we may consider feature selection criterion. Here, we develop an entropy-based selection as follows. We rely on the conditional entropy $H(y|x)$ to evaluate the amount of uncertainty remaining about y if x is known. For a given HR grid point p, we select the LR grid points with the lowest conditional entropy values of wind y at point p knowing field x at a LR grid point q:

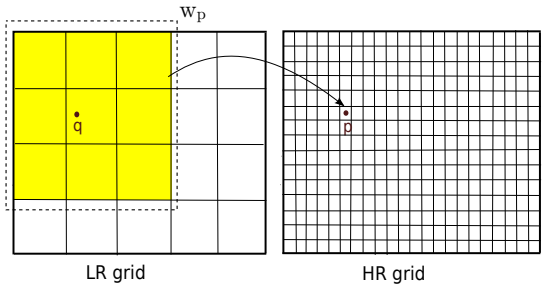


Fig. 3: **Local information scheme**: the 9 grid points in the local neighborhood defined by the yellow box around the LR grid point q are used to define regression variables.

$$H(y_p|x_q) = -\sum_{m=1}^{M}\sum_{n=1}^{n} P(y_p = y_m, x_q = x_n)\cdot \log P(y_p = y_m|x_q = x_n) \quad (11)$$

where $P(y_p = y_j, x_q = x_i)$ and $P(y_p = y_j|x_q = x_i)$ are the joint probability and the conditional probability respectively of y and x. We consider here a discretized setting. Indices m (resp. n) refer to the number M (resp. N) of discrete states of random variables $\{y_p\}$ (resp. $\{x_p\}$).

As an example, Figure 4 reports conditional entropy values for a coastal feature HR grid point (red squares with black face). For the computation of the conditional entropy values, the HR winds are discretized as {strong, medium, weak}. Thus, the maximum entropy for our study is $\log_2(3)$ (≈ 1.59) and the minimum is 0 which means no uncertainty. The selected 9 LR grid points with the lowest conditional entropy are indicated by the red circles. Interestingly, in the reported example, the selected LR grid points do not resort to the LR neighboring window. It means that not all LR neighboring points provide the same amount of information to predict the HR information. This is particularly important for coastal grid points where the differences between HR and LR data are higher than in the offshore area.

E. Learning and implementation issues

In the considered downscaling setting, we learn a different regression model at each HR grid point. Given such point-specific regression models, the reconstruction of HR SAR wind fields given a LR model prediction x comes to applying the trained regression functions to each HR grid point.

Regarding training issues, at each HR grid point, a k-fold cross-validation (k=19) is performed for calibration and evaluation issues. The whole samples of the SAR-ECMWF pairs are randomly partitioned into 19 folds subsamples. A single subsample is retained as the validation set while the remaining data are used for training. The cross-validation process is repeated 19 times (the folds), with each of about five percent subsamples used exactly once as the validation set. We evaluate the regression performance (for instance, the mean regression error) for validation datasets.

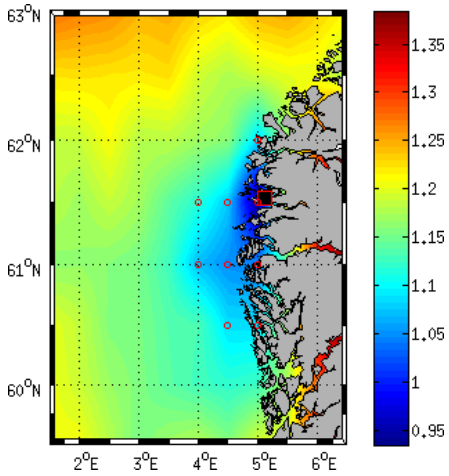


Fig. 4: **Conditional entropy values (Eq.11) for a coastal HR grid point**: the red square with black face shows the HR grid point p and the red circles indicate the 9 LR grid points with the lowest conditional entropy values (Eq.11). The conditional entropy varies between 0 and 1.59 ($\log_2(3)$) in our case. High entropy corresponds to high uncertainty.

IV. Results

We carry out a qualitative and quantitative evaluation of the proposed approaches based on the considered SAR-ECMWF dataset (Section II). We compare four regression methods, namely Nearest Neighbor, K-Nearest Neighbors, Multiple Linear Regression (MLR) and Support Vector Regression (SVR) methods. We also evaluate the combination of these regression models to three types of regression variables definition as described in the previous section. For the SVR, we use a radial basis function as kernel function. Empirical results demonstrated the relevance of this choice compared to other kernels.

As the study area involves different situations, we analyze the regression performance for twelve grid points (Figure 5), which account for offshore, coastal and within-fjord conditions.

A. Comparison of the downscaling models

We first evaluate the influence of the information type on downscaling performance. Figure 6 reports the mean regression error in $\mathrm{m\,s^{-1}}$ for three types of regression variables (cf. Section III-D): global information, local information and entropy-based information. Both MLR (Figure 6a) and SVR methods (Figure 6b) achieve much better downscaling performance using local and entropy-based information than global information for the selected HR grid points (cf. Figure 5), with a gain around $0.5\,\mathrm{m\,s^{-1}}$. The performance of local and entropy-based information is very close. However, the entropy-based information provides visually more consistent HR fields to avoid "tiling effects" [36].

For local and entropy-based information, we analyzed the sensitivity to the selected number of LR grid points. Figure 7 shows the influence of the number of LR grid points used in the regression model on the downscaling performance at

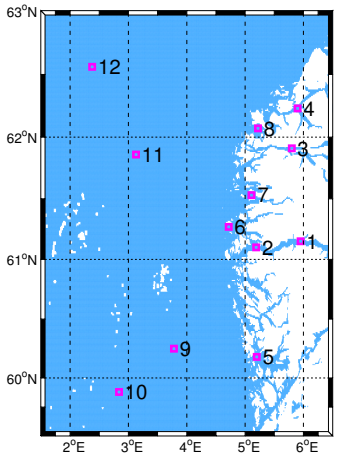
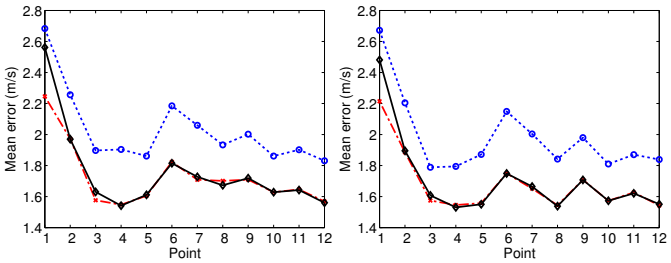


Fig. 5: **Study area and points selected for the evaluation of regression error statistics**: the study area is southwest coast sea (in blue) of Bergen described in Section II. The grid points 1 to 4 account for fjord, 5 to 8 for coastal, and 9 to 12 for offshore condition evaluations.



(a) MLR (b) SVR

Fig. 6: **Comparison of the three types of regression variables**: global information (blue circles, dashdots); local information (red x-marks, dashdot lines) and entropy-based information (black diamonds, solid lines). We report the mean regression error in $\mathrm{m\,s^{-1}}$ for different regression models: Multiple Linear Regression (a) and Support Vector Regression (b). The study area is located to the west of Bergen in the Norwegian Sea. The data used are described in Section II. We proceed to cross-validation experiments (cf. Section III-E) to evaluate regression error statistics.

HR grid point 8 ((N62.07°, E5.22°) in Figure 5). 9 LR grid points give the lowest prediction error for the analog and MLR method. The SVR models achieve the best performance using between 9 and 25 grid points. A similar pattern has been observed at other HR grid points. Overall, 9 LR points were selected for the subsequent analysis.

The comparison of the different regression models clearly stresses (Figure 8) that the SVR model, with both local and entropy-based information, outperforms analog and MLR models. It achieves a mean regression error around $1.7\,\mathrm{m\,s^{-1}}$. The errors for grid points 1 and 2 within the fjord are significantly higher than for the other points for all approaches. The same conclusion can be drawn from the analysis of correlation coefficients and quantiles (not illustrated here).

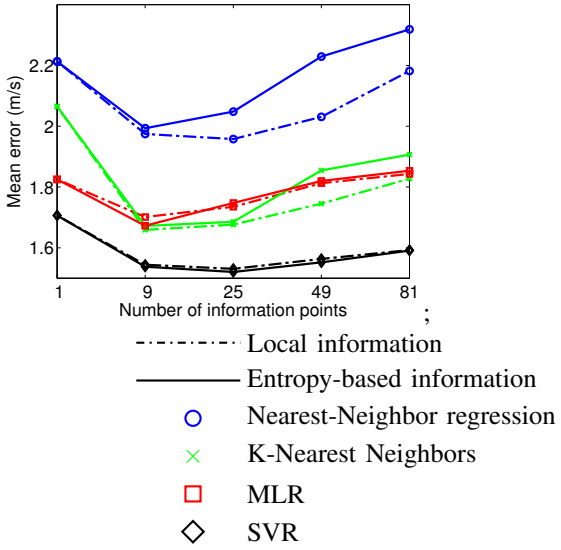


Fig. 7: **Influence of the number of regression variables (x-axis) on the prediction error (y-axis) at grid point 8 ((N62.07°, E5.22°) in Figure 5)**: Nearest-Neighbor regression (blue circles), K-Nearest Neighbors (green x-marks), Multiple Linear Regression (MLR, red squares) and Support Vector Regression (SVR, black diamonds). For each grid point and for each regression method, we compare two types of regressions variables, namely local information (dashdot lines) and entropy-based information (solid lines) (see Section III for details).

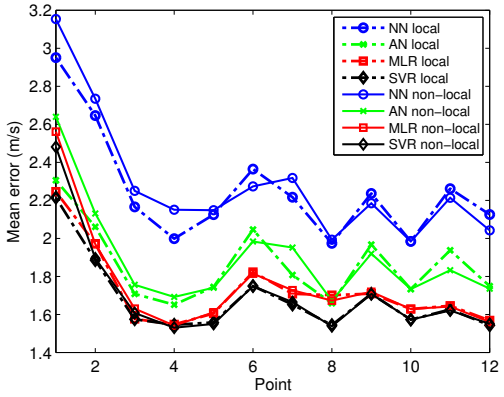


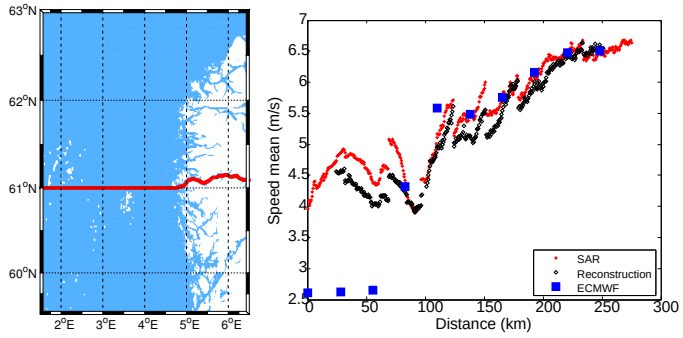
Fig. 8: **Mean downscaling error in $\mathrm{m\,s^{-1}}$ (y-axis) at the different HR grid points (x-axis, see Figure 5 for their locations)** for different approaches: Nearest-Neighbor model (NN, blue circles), K-Nearest Neighbor model (AN, green x-marks), Multiple Linear Regression (MLR, red squares) and Support Vector Regression (SVR, black diamonds). For each method, we compare downscaling performance using local information (dashdot lines) and entropy-based information (solid lines) (see Section III for details).

Overall, this experimental evaluation demonstrates the relevance of the SVR model with 9 LR grid points (using local or entropy-based information) compared to the other models (global information, linear regression and analog regression).

B. Analysis of the HR wind fields emulated by the SVR-based model

In addition to the quantitative comparison reported above, we further analyze the relevance of the SVR-based emulation of HR wind fields. The emulation area is limited to E2.0°– E6.0° and N60.0°– N62.5° to avoid border effects.

Figure 9 shows the evolution of the mean eastern wind of the ECMWF (blue squares), SAR (red stars) and downscaled (black diamonds) fields as a function of the distance from the easternmost point (N61.09°, E6.50°, Figure 9a). At each grid point, the mean eastern wind is computed as the mean magnitude of all wind data with a wind direction between $-45°$ and $45°$. As illustrated by the SAR wind data, the presence of chain of mountains parallel to the coast creates wave patterns in eastern wind conditions. The ECMWF fields do not retrieve such patterns. By contrast, the downscaled winds involve a wave pattern very similar to that of SAR winds. In particular, the wavelength of the wave patterns is well reconstructed by the SVR model.



(a) Transect line (b) Distance from the east coast

Fig. 9: **Mean eastern wind along a transect perpendicular to the coast (a, red line) in** $\mathrm{m\,s^{-1}}$: ECMWF (blue squares), SAR (red stars) and downscaled (black diamonds) fields. The x-axis refers to the distance from the easternmost point (N61.09°, E6.50°) in Figure 9a towards offshore. The mean eastern wind is computed at each grid point as the mean wind magnitude of all wind data with a wind direction between $-45°$ and $45°$.

Figure 10 compares the distributions of the ECMWF (left), SAR (middle) and downscaled (right) winds at fjord point 4 (N62.23°, E5.90°) (cf. Figure 5). ECMWF and SAR distributions show significant differences both in terms of wind direction and magnitude, which can be interpreted as resulting from local topography effects. By contrast, the SVR model succeeds in downscaling wind patterns similar to the SAR data.

We report wind speed scatterplots for grid points 4 (N62.23°, E5.90°) and 8 (N62.07°, E5.22°) to further analyse the differences of the ECMWF, SAR and downscaled wind data (Figure 11). Interestingly, due to their local topographic configurations, these two examples involve an underestimation and overestimation of HR wind speeds by the ECMWF data. The SVR-based downscaling retrieves consistent HR wind speeds, even if the ECMWF data poorly match the HR pattern.

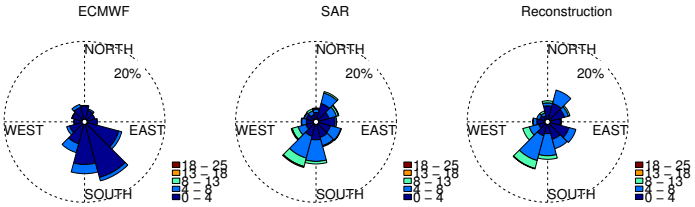
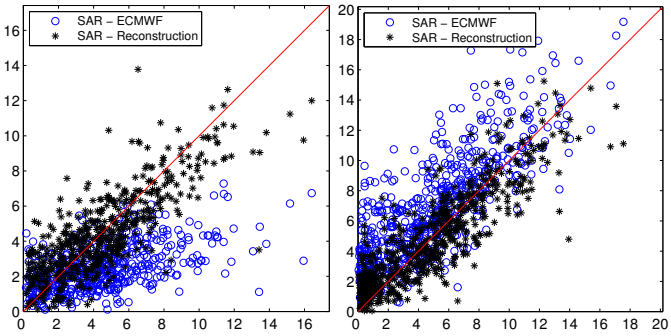


Fig. 10: **Distribution of the wind data (direction and speed) at point 4 (N62.23°, E5.90°) in Figure 5**: ECMWF data (left), SAR data (middle) and downscaled data (right). We compute the wind roses as in Figure 2.



(a) Point 4 – (N62.23°, E5.90°) (b) Point 8 – (N62.07°, E5.22°)

Fig. 11: **Wind speed scatterplots between ECMWF, SAR and downscaled wind data for grid points 4 (N62.23°, E5.90°) and 8 (N62.07°, E5.22°)**: ECMWF vs. SAR (blue circles), downscaled vs. SAR (black stars). x-axis denotes SAR winds when y-axis denotes ECMWF or downscaled winds. Wind speed is in $\mathrm{m\,s^{-1}}$. The red lines represent a perfect match. Regarding downscaled vs. SAR scatterplot, we exploit the downscaled wind fields generated within the k-fold cross-validation procedure for the randomly generated validation datasets (see Section III.E for details).

This is regarded as a direct outcome of the flexibility of the non-linear SVR learning.

Besides, we report two examples of downscaled HR SAR wind fields in Figure 12. There is no emulation for the white points that match the oil platform locations where SAR wind measurements are erroneous. The 2009-07-18 situation (Figure 12a) corresponds to a north-east wind. Overall, the weak wind behind the coastline and the strong wind following the dominant wind direction are well reconstructed, except some texture-like turbulence patterns. Such patterns are rather random-like patterns, which may hardly be captured by a deterministic downscaling model. The 2006-02-05 situation (Figure 12b) shows a western wind. Whereas the wind slows down close to the coast, it accelerates towards the north of the coastal zone. The strong-weak-strong shift is well reconstructed by the SVR model.

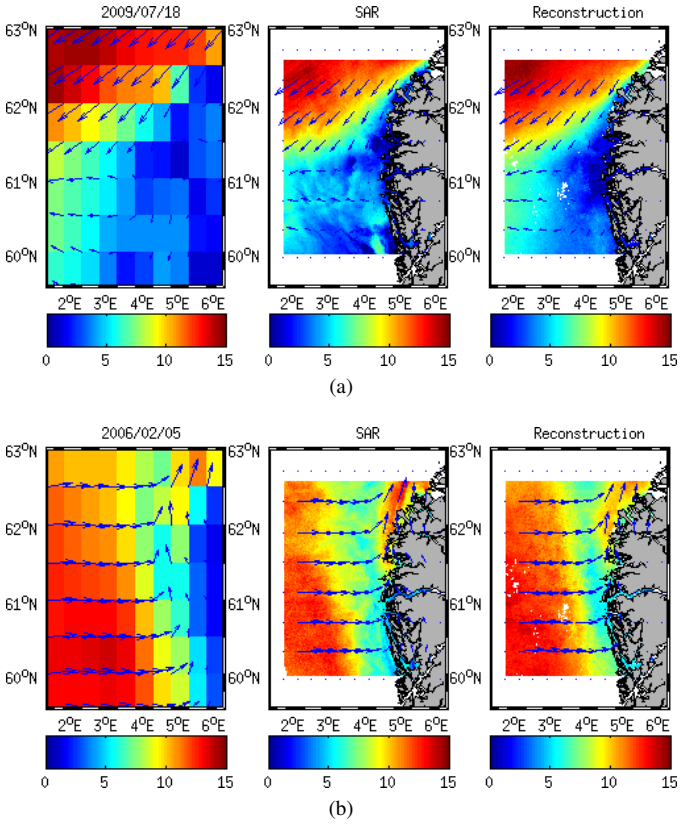


Fig. 12: **Examples of downscaled wind fields for two dates, 2009-07-18 (top) and 2006-02-05 (bottom)**: from left to right, ECMWF field, SAR field, downscaled field. For the first example, the ECMWF data are delivered at 12h UTC and the SAR data are acquired at 10h09 UTC. For the second example, the ECMWF data are delivered at 00h UTC 2006-02-06 and the SAR data are acquired at 21h05 UTC. There is no emulation for the white points that corresponds to oil platform locations where SAR wind measurements are erroneous. The color and the arrows indicate the wind speed in $\mathrm{m\,s^{-1}}$ and the wind direction.

V. Conclusion and Future Work

In this study, we have shown that learning-based downscaling models, especially SVR models, provide a relevant solution for the emulation of HR wind fields from operational ECMWF data. This learning-based strategy primarily relies on the construction of a dataset of co-located LR ECMWF and HR SAR-derived wind fields. The quality and the representativeness of this dataset are obviously critical issues. Using a simple filtering scheme, based on the evaluation of the mean square difference between the LR and HR fields in the offshore region, we were able to detect poor matches between the ECMWF and SAR fields. Such situations, which typically relate to forecasting uncertainty as well as to erroneous SAR-derived measurements (*e.g.,* due to heavy rainfalls), were withdrawn from our analysis and did not affect the learning of the downscaling models. This further supports that the evaluation of the impact of the inaccuracy of the low-resolution condition should be considered. In the future work, this could be discussed by including possible combination to ensemble forecasts.

Overall, our dataset comprised 758 validated pairs of ECMWF-SAR data, *i.e.,* $\sim 90\%$ of the initial dataset. To make feasible the implementation of the proposed learning-based scheme with k-fold cross-validation, we exploited a cluster-based implementation (a cluster of 70 servers with two processors of 12 cores). It resorted to a total execution time 1600 times smaller than a single-server implementation, *i.e.,* about 6 hours to be compared to 400 days. Overall, we reached a mean precision of $1.6\,\mathrm{m\,s^{-1}}$ offshore, $1.8\,\mathrm{m\,s^{-1}}$ in coastal areas and $2.6\,\mathrm{m\,s^{-1}}$ in fjord regions. While these results are regarded as a first validation of the proposed SVR framework, we expect that appending new ECMWF-SAR data will improve reconstruction performance, especially in fjord areas. Future work should further investigate and evaluate, with respect to in situ wind data, downscaling performance in relation to size of the training dataset and ECMWF-SAR consistency check.

From a methodological point of view, the main conclusions drawn from this study are two-fold. SVR-based models have been shown to outperform the popular linear and analog regression models. Beyond the improvement of downscaling performance, we have stressed that SVR can be regarded as a generalization of the later regression models. Regarding the computational complexity of the SVR model, one should distinguish two aspects: the training step and the regression step. For the training step, we exploited a cluster-based implementation, which made feasible the calibration of SVR model parameters for point-specific models in a reasonable time (about half an hour for the considered 250×400 HR grid points). For the regression step, the computational complexity of the SVR resorts to a sum over training samples, similarly to the analog regression. However, from the SVR theory, one can expect many SVR weights (Eq.10) to be null. In our case study, offshore, coastal and fjord zones involved respectively about 10%, 25% and 50% of the training samples. Hence, the SVR-based reconstruction involved a significant reduction of the computational complexity compared to the analog regression. Future work will further explore SVR-based models for downscaling applications with a specific emphasis on the definition and selection of regression variables, including non-local and coupled global-local variables, as well as on the extraction of spatially-sparser SVR-based models.

Our learning-based approach could be applied to any kind of couple of low-resolution regular time-step ant high-resolution irregular time sampling data sets. A nice feature of ECMWF wind fields is that we may use a 40-year-long time series (ERA40) outputs with a worldwide coverage and then the approach can be tested in any regions where SAR data are available. Beyond sea surface winds, the genericity of the proposed SVR-based models also advocate applications to the downscaling of other sea surface parameters (e.g., sea surface temperature, ocean color), which also involve irregular space-time sampling.

REFERENCES

[1] H. Von Storch, B. Hewitson, and L. Mearns, "Review of empirical downscaling techniques," in *RegClim Spring Meeting*, 2000.

[2] E. Zorita and H. Von Storch, "The analog method as a simple statistical downscaling technique: comparison with more complicated methods," *Journal of Climate*, vol. 12, no. 8, pp. 2474–2489, 1999.

[3] F. Monaldo, V. Kerbaol, P. Clemente-Colón *et al.*, "The SAR measurement of ocean surface winds: an overview," in *Proceedings of the Second Workshop Coastal and Marine Applications of SAR*, 2003, pp. 2–12.

[4] V. Kerbaol, "Improved bayesian wind vector retrieval scheme using ENVISAT ASAR data: principles and validation results," in *Proceedings of ENVISAT Symposium*, 2007, pp. 23–27.

[5] R. Wilby, S. Charles, E. Zorita, B. Timbal, P. Whetton, and L. Mearns, "Guidelines for use of climate scenarios developed from statistical downscaling methods," *IPCC TGCIA*, 2004.

[6] J. H. Christensen, B. Hewitson, A. Busuioc, A. Chen, X. Gao, R. Held, R. Jones, R. K. Kolli, W. Kwon, R. Laprise *et al.*, *Regional climate projections*. Cambridge University Press, 2007, ch. 11, pp. 847–940.

[7] R. E. Benestad, I. Hanssen-Bauer, and D. Chen, *Empirical-statistical downscaling*. World Scientific Pub Co Inc, 2008.

[8] A. Busuioc, H. Von Storch, and R. Schnur, "Verification of gcm-generated regional seasonal precipitation for current climate and of statistical downscaling estimates under changing climate conditions." *Journal of Climate*, vol. 12, no. 1, 1999.

[9] B. Tang, W. W. Hsieh, A. H. Monahan, and F. T. Tangang, "Skill comparisons between neural networks and canonical correlation analysis in predicting the equatorial pacific sea surface temperatures." *Journal of Climate*, vol. 13, no. 1, 2000.

[10] A. Wu, W. W. Hsieh, and B. Tang, "Neural network forecasts of the tropical pacific sea surface temperatures," *Neural Networks*, vol. 19, no. 2, pp. 145–154, 2006.

[11] V. Vapnik, S. E. Golowich, and A. Smola, "Support vector method for function approximation, regression estimation, and signal processing," *Advances in neural information processing systems*, pp. 281–287, 1997.

[12] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] G. Dreyfus, J.-M. Martinez, M. Samuelides, M. B. Gordon, F. Badran, and S. Thiria, *Apprentissage statistique: Réseaux de neurones-Cartes topologiques-Machines à vecteurs supports*. Editions Eyrolles, 2011.

[14] L. HE, "Émulation statistique de champs de vent haute rsolution," Ph.D. dissertation, Télécom Bretagne with l'École Doctorale Sicma, Brest, France, 2014.

[15] J. W. Kidson and C. S. Thompson, "A comparison of statistical and model-based downscaling techniques for estimating local climate variations." *Journal of Climate*, vol. 11, no. 4, 1998.

[16] S. Aguilar-Martinez and W. W. Hsieh, "Forecasts of tropical pacific sea surface temperatures by neural networks and support vector regression," *International Journal of Oceanography*, vol. 2009, 2009.

[17] K. Goubanova, V. Echevin, B. Dewitte, F. Codron, K. Takahashi, P. Terray, and M. Vrac, "Statistical downscaling of sea-surface wind over the peruchile upwelling region: diagnosing the impact of climate change from the ipsl-cm4 model," *Climate Dynamics*, pp. 1–14, 2010.

[18] A. Stoffelen and D. Anderson, "Scatterometer data interpretation: Estimation and validation of the transfer function CMOD4," *Journal of Geophysical Research*, vol. 102, no. C3, pp. 5767–5780, 1997.

[19] H. Hersbach, "Comparison of C-band scatterometer CMOD5.N equivalent neural winds with ECMWF," *J. Atmos. Oceanic Technol.*, vol. 27, pp. 721–736, 2010.

[20] B. Zhang and W. Perrie, "Cross-polarized synthetic aperture radar: A new potential measurement technique for hurricanes," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, pp. 531–541, 2012.

[21] F. M. Monaldo, D. R. Thompson, R. C. Beal, W. G. Pichel, and P. Clemente-Colón, "Comparison of sar-derived wind speed with model predictions and ocean buoy measurements," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 39, no. 12, pp. 2587–2600, 2001.

[22] B. R. Furevik and E. Korsbakken, "Comparison of derived wind speed from synthetic aperture radar and scatterometer during the ers tandem phase," *Geoscience and Remote Sensing, IEEE Transactions on*, vol. 38, no. 2, pp. 1113–1121, 2000.

[23] I. Orlanski, "A rational subdivision of scales for atmospheric processes," vol. 56, no. 5, pp. 527–530, May 1975.

[24] R. Mayenon, *Météorologie marine*. Éditions maritimes et d'outre-mer, 1982.

[25] P. Beaucage, A. Glazer, J. Choisnard, W. Yu, M. Bernier, R. Benoit, and G. Lafrance, "Wind assessment in a coastal environment using synthetic aperture radar satellite imagery and a numerical weather prediction model," *Canadian Journal of Remote Sensing*, p. 368377, 2007.

[26] W. Koch and F. Feser, "Relationship between sar-derived wind vectors and wind at 10-m height represented by a mesoscale model." *Monthly weather review*, vol. 134, no. 5, 2006.

[27] K.-F. Dagestad, H. Morten W., J. Johnny A. *et al.*, "Development and validation of a sar wind emulator," Nansen Environmental and Remote Sensing Center, Tech. Rep., 2009.

[28] J. Walmsley, R. Barthelmie, and W. Burrows, "The statistical prediction of offshore winds from land-based data for wind-energy applications," *Boundary-Layer Meteorology*, vol. 101, pp. 409–433, 2001.

[29] M. Minvielle, C. Cassou, R. Bourdallé-Badie, L. Terray, and J. Najac, "A statistical–dynamical scheme for reconstructing ocean forcing in the atlantic. part II: methodology, validation and application to high-resolution ocean models," *Climate dynamics*, vol. 36, no. 3, pp. 401–417, 2011.

[30] E. N. Lorenz, "Atmospheric predictability as revealed by naturally occurring analogues," *Journal of the Atmospheric sciences*, vol. 26, no. 4, pp. 636–646, 1969.

[31] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[32] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.

[33] V. Vapnik and A. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

[34] B. Schölkopf and A. Smola, *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[35] V. Cherkassky and Y. Ma, "Practical selection of SVM parameters and noise estimation for SVM regression," *Neural networks*, vol. 17, no. 1, pp. 113–126, 2004.

[36] L. He, R. Fablet, B. Chapron, and J. Tournadre, "Statistical emulation of high resolution SAR wind fields from low-resolution model predictions," in *IGARSS 2014 : IEEE International Geoscience and Remote Sensing Symposium*, 2014.

Liyun He graduated from "Beijing University of Posts and Telecommunication" (BUPT), China and "École Nationale Supérieure des Télécommunications de Bretagne" (TELECOM Bretagne), France, in 2006. She received her Ph.D. degree in Information and Communication Science and Technology from University of Western Brittany, France, in 2014. She currently holds a post-doctoral position in TELECOM Bretagne. Her main research interests include advanced statistical methods for télécommunications and oceanography applications.

Ronan Fablet graduated from "École Nationale Supérieure de l'Aéronautique et de l'Espace (SUPAERO)", France, in 1997. He received the Ph.D. degree in Signal Processing and Telecommunications from the University of Rennes, France, in 2001. In 2002, he was an INRIA post-doctoral fellow at Brown University, RI, USA. From 2003 to 2007, he held a full-time research position at Ifremer Brest in the field of signal and image processing applied to fisheries science. In 2008 he joined the signal and communications department of Telecom Bretagne as an Associate Professor, and has been holding a Professor position since 2012. In 2011, he was a visiting researcher at IRD/IMARPE, Peru (Peruvian Sea Research Institute). His main interests include statistical methods for signal processing and computer vision and applications to ocean remote sensing.

Bertrand Chapron received the B.Eng. degree from the Insti- tut National Polytechnique de Grenoble, Grenoble, France, in 1984 and the Ph.D. degree in fluid mechanics from the University of Aix-Marseille II, Marseille, France, in 1988. He spent three years as a Post-Doctoral Re- search Associate at the NASA/GSFC/Wallops Flight Facility, Wallops Island, VA. He has experience in applied mathematics, physical oceanography, electromagnetic waves theory, and its application to ocean remote sensing. He is currently responsible for the Oceanography from Space Laboratory, IFREMER, Plouzan, France.

Jean Tournadre graduated from Ecole Centrale de Lyon France in 1981, received a Ph.D. in Geophysics from University of Clermont II in 1984, then a HDR on physical remote sensing methods from University of Paris 7 in 1998. He did a post doctoral research fellowship at Scripps Institution of Oceanography, University de California San Diego, US, from 1984 to 1986 then was visiting scientist at NCAR, Boulder, US, in 1987. He finally got a permanent position at the Satellite Oceanography department of IFREMER, with main expertise on altimetric data processing for ocean wind, wave, current and icebergs observation.