IEEE International Symposium on Geoscience and Remote Sensing IGARSS
2014, Pages 3914-3917
https://doi.org/10.1109/IGARSS.2014.6947340
https://archimer.ifremer.fr/doc/00256/36733/

Statistical emulation of high-resolution sar wind fields from low-resolution model predictions

He Liyun [1], Chapron Bertrand [1], Tournadre Jean [1], Fablet Ronan [2]

[1] IFREMER, Lab Oceanog Spatiale, Brest, France.
[2] Telecom Bretagne, Departement SC Brest, FRANCE

Abstract :

This paper addresses the reconstruction of high-resolution (HR) sea surface wind fields (typically, at a spatial resolution of 1 k m). The availability of such HR fields is critical for numerous issues, e.g. coastal management, offshore structures, oil spill disaster tracking, etc. Satellites, especially from Synthetic Aperture Radar (SAR) systems, can monitor the ocean surface at a spatial resolution of a few meters. SAR wind fields are operationally produced with spatial resolutions of less than 1 k m [1, 2]. 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. By contrast, model predictions, such as European Center for Medium-range Weather Forecast (ECMWF) wind fields, are typically delivered with a high temporal resolution (e.g. every 3 h or 6 h), but with a low spatial resolution (similar to 50 km x 50 km). The question of the combination of numerical model predictions and SAR wind fields naturally arises to deliver HR wind fields at sea surface anywhere and anytime. Here, we state this issue as the statistical learning of transfer functions between low-resolution (LR) model predictions and the associated HR SAR fields. We investigate the extent to which such regression functions can be learnt from a set of co-located HR and LR fields. Both local and non-local schemes as well as linear and non-linear regression methods are considered. As a case-study, we carry out numerical experiments for a coastal area off Norway, which involves complex LR-to-HR situations.

Keywords : Statistical downscaling, High resolution, Support Vector Regression (SVR), SAR coastal wind

1. INTRODUCTION

The derivation of local scale information from integrations of coarse-resolution general circulation models with the help of statistical models fitted to present observations is generally referred to as statistical downscaling [3, 4]. Statistical downscaling offers an alternative solution to overcome the scale mismatch of both numerical model predictions and satellite observations. It is becoming popular due to its relative simplicity and low computing costs [3]. Statistical downscaling is particularly useful for heterogeneous environments with complex geography such as strong environmental gradients due to the presence of an island, a mountain or in a continent/ocean context where physical processes are difficult to model directly [5]. For this kind of configuration, [6] shows that statistical downscaling provides a pragmatic approach to model local parameters from large-scale climate information.

In this study, we display our interest in regression-like techniques using linear or non-linear formulations. It seems that more complicated regression techniques do not significantly improve the quality of prediction [4, 7, 8]. As most statistical downscaling models found in the literature use global information as regression variables, it means that other ways of exploiting LR information should be considered simultaneously. The global information is represented by the projection coefficients, called Principle Components (PCs), of large-scale fields in the space spanned by the leading observed Empirical Orthogonal Functions (EOFs) [9, 4, 7, 8, 10]. The use of global information may lead to the loss of information accuracy. In this study, we propose two more local information schemas: simple local information and Entropy-based information, as HR variability may depend directly on the more local variables.

This paper is organized as follows: Section 2 briefly presents the considered data and the study area. The proposed learning-based approach is described in Section 3. Section 4 reports and discusses numerical experiments.

2. DATA AND STUDY AREA

The LR data used in this work are the ECMWF analysis data at 0.5° spatial resolution. These data are available every $6\,\mathrm{h}$. The HR SAR data, issued from the ENVironmental SATellite (ENVISAT) and processed by the Collecte Localisation Satellites (CLS) centre, achieve a spatial resolution of 0.01°.The study area is the Southwestern coastal sea of Bergen, where the sea wind situations are very complex due to the topography, especially the presence of islands and fjords. This makes this area an interesting study area. Statistical analyses show that the two data sets are very similar at LR in the offshore

area and more and more different while getting closer to the coast. The distributions of wind directions and speeds also vary a lot spatially. These features motivate the design of point-specific transfer functions.

Overall, we consider a dataset of 758 pairs of ECMWF and SAR data available in the area $59°50' - 63°0'$ in latitude and $1°50' - 6°50'$ in longitude. Here, the ENVISAT SAR data were acquired from 2005 to 2010. Each SAR data is co-located with the temporally closest EMWF field, thus they may have a maximum difference of 3 h. Pairs of ECMWF and SAR data that are very different are rejected.

3. 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 low resolution variables. The problem is modeled as

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

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

3.1. Learning scheme

Our goal is to learn regression function f from a training set $\{x_k, y_k\}$ of co-located HR and LR wind fields. It resorts to retrieving the optimal regression function $f^* \in \mathcal{F}$ which minimizes the regression errors $y - f^*(x)$, with $\mathcal{F}$ a set of regression function. Generally, the analog and multiple linear regression (MLR) methods are used as downscaling schemes [4]. Here, we investigate an optimal non-linear kernel-based regression model, namely Support Vector Regression (SVR). It can be regarded as a linear regression model in a space defined by a non-linear mapping function Φ [11]:

$$f(x, \omega) = \omega^t \Phi(x) + b \tag{2}$$

where ω is the regression model weight vector. In ε-SVR regression, the regression model weights are obtained by resolving the optimization problem:

$$\frac{1}{2}\|\omega\|^2 + \mathrm{C} \sum_{i=1}^{n} (\xi_i + \xi_i^*) \tag{3}$$

constrained to:

$$\begin{cases} y_i - \langle \omega, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle \omega, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \varepsilon, \xi_i, \xi_i^* \geq 0 \end{cases} \qquad \forall i \in (1, n) \tag{4}$$

where ε is the accuracy term, ignoring errors between the observed and predicted value of y smaller than ε. Slack variables ξ_i and ξ_i^* are error measurements above and below the ε-insensitivity zone, respectively. Regularization parameter C determines the trade-off between the flatness of f and the amount up to which deviations larger than ε are tolerated [12].

Then the regression model can be rewritten according to the kernel function K, defined by $K(x, x') = \langle \Phi(x), \Phi(x') \rangle$:

$$f(x, \omega) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b \tag{5}$$

where $\{(x_i, y_i)\}$ refers to the available training data, and $(\alpha_i - \alpha_i^*)$ sets the relative weight of each training data in the regression model. Given a kernel model, the training of the SVR model resorts to the inference of the optimal weight vector according to a margin-based criterion. We let the reader refer to [13] for further details.

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

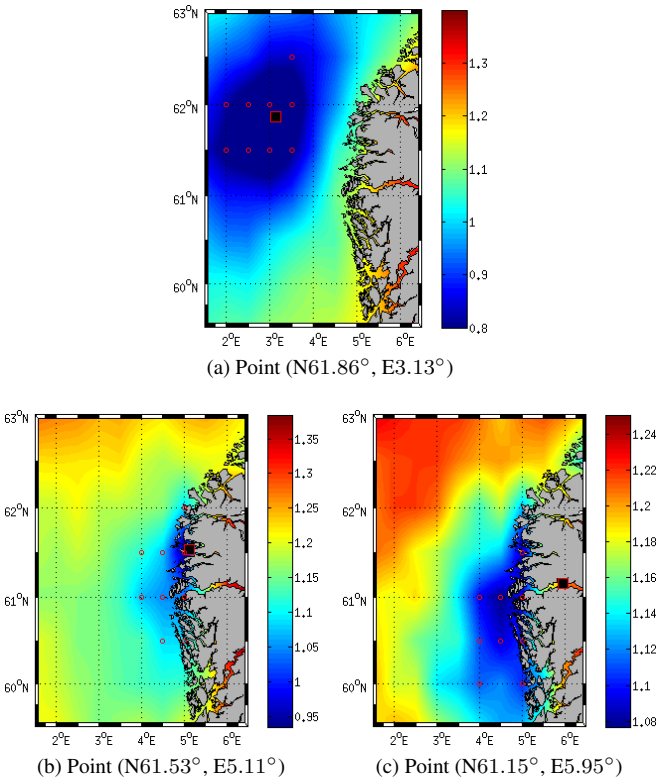
3.2. Selection of the regression variables

Different approaches may be considered to select the regression variables [14]. Here we investigate two different approaches: a simple approach which exploits the low resolution wind information within a local neighborhood around the HR grid point and a more sophisticated approach which involves a prior selection step and selects the relevant LR information to achieve the prediction at a given HR grid point. For the second approach, we propose an original entropy-based scheme for the selection step. For a given HR grid point p, we select the LR grid points with the lowest values of the conditional entropy of field y at point p knowing field x at a LR grid point q:

$$H(y_p|x_q) = -\sum_{j=1}^{m}\sum_{i=1}^{n} P(y_p = y_j, x_q = x_i) \log P(y_p = y_j | x_q = x_i) \tag{6}$$

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 for each joint case. $H(y_p|x_q)$ is a measure of amount of uncertainty remaining about y_p after x_q is known. For each HR grid point, we select k LR grid points with the lowest conditional entropy (lowest uncertainty).

Figure 1 shows conditional entropy for an offshore (a), coastal (b) and fjord (c) feature HR grid point (red squares). The selected 9 LR grid points with the lowest conditional entropy are indicated by the red circles. We note that the selected LR grid points are adjacent to the given offshore feature HR grid point, in comparison to a given coastal or fjord HR grid point where they are more dispersed.



(a) Point (N61.86°, E3.13°)

(b) Point (N61.53°, E5.11°)

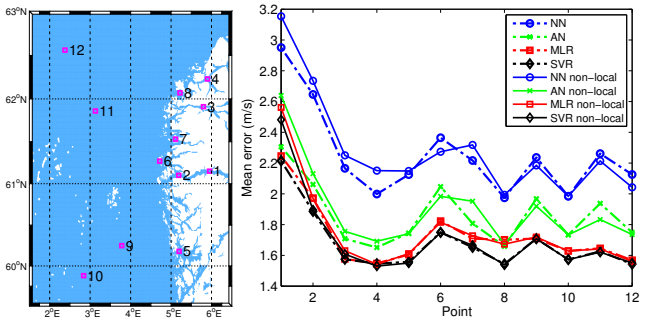
(c) Point (N61.15°, E5.95°)

Fig. 1: Conditional entropy for an offshore (a), coastal (b) and fjord (c) feature HR grid point. In each panel, The red square shows the HR grid point p and the red circles indicate the 9 LR grid points with the lowest conditional entropy.

4. RESULTS

We carry out a qualitative and quantitative evaluation of the proposed approaches based on the considered SAR-ECMWF dataset. Three other approaches such as Nearest Neighbor (NN), Weighted Average ANalog(AN) and Multiple Linear Regression (MLR) methods [4] are evaluated in our experiments.

We first proceed to cross-validation experiments to evaluate regression error statistics. Ninety five percent of the SAR-ECMWF pairs are randomly sampled for training and the remaining pairs are used to evaluate regression error. Nine LR grid points are used for the two variables selection approaches. As the study area involves different situations, we analyze the regression performance for twelve different grid points (Figure 2a). These points account for offshore, coastal and within-fjord downscaling features. Overall, as stressed by Figure 2b, the SVR model, with both patch-based regression (referred to as local) and non-local entropy-based selection of the regression variables (§ 3.2), outperforms the other models and achieves a mean regression error around $1.7\,\mathrm{m\,s^{-1}}$. The errors for grid point 1 and 2 within the fjord are significantly higher than for the other points for all approaches. Similar results which are not illustrated here, such as the correlation coefficient and quantiles, show the same tendency as of the mean error.



(a) (b)

Fig. 2: Study area and points selected for the evaluation of regression error statistics (a); mean regression error in $\mathrm{m\,s^{-1}}$ for different approaches, namely: Nearest-Neighbor regression (NN, blue circles), ANalog regression (AN, green x-marks), Multiple Linear Regression (MLR, red squares) and Support Vector Regression (SVR, black diamonds). For each regression method, we compare a patch-based regression, referred to as local (dashdot lines) to a non-local entropy-based selection of the regression variables (solid lines) (b).

Figure 3 shows an example of the reconstructions of a HR SAR wind field. The reconstructed case is that of 2009-03-04 where the wind is very strong. For each HR grid point, 9 LR grid points (k value, § 3.2) with the lowest conditional entropy are selected for the learning. There is no reconstruction for the white points that match the oil platform locations or the points where there is no similar LR situation for an analog approach. In this reconstructed case, both MLR and SVR show good results. Compared to the use of a local neighborhood (Figure 4), a non-local entropy based selection of the regression variables is particularly appropriate to avoid the "tiling effect".

5. CONCLUSION

In this study, we show that machine learning models based on non-linear Support Vector Regression (SVR) method, combined either with local or non-local information, perform better than more classical models, *e.g.* those based on analog method or Multiple Linear Regression (MLR). The SVR based models produce HR variability very close to the reference local variability observed by SAR, especially in coastal area. Furthermore, the reconstructed wind fields preserve the statistical distribution properties of SAR wind fields.

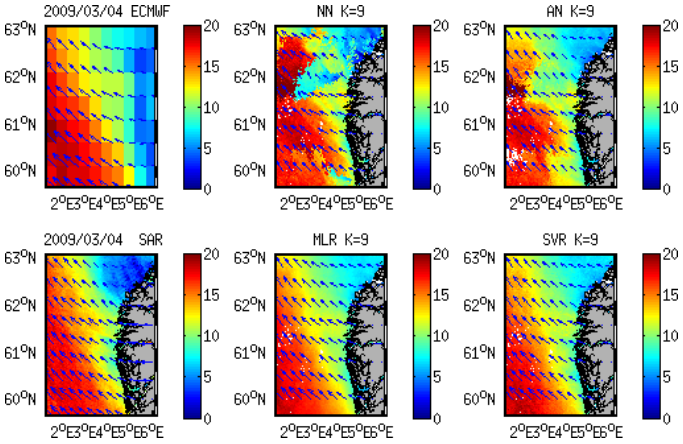


Fig. 3: Reconstructed wind fields for ECMWF data of 2009-03-04 (top left) with different learning methods using non-local entropy-based selection: Nearest-Neighbor regression (NN, top middle), ANalog regression (AN, top right), Multiple Linear Regression (MLR, bottom middle) and Support Vector Regression (SVR, bottom right). The corresponding SAR data (bottom left) is used as reference.

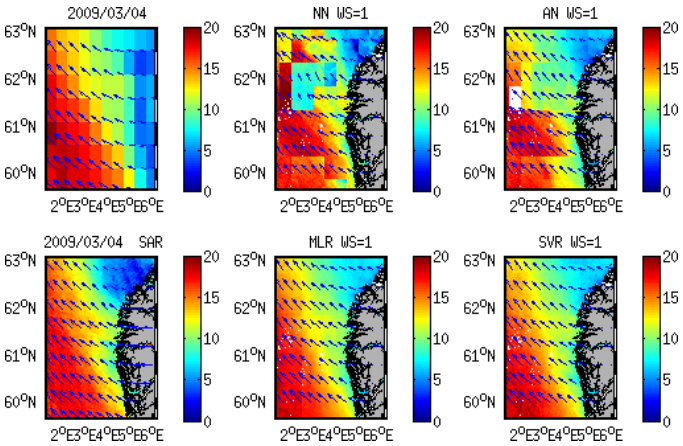


Fig. 4: Reconstructed wind fields for ECMWF data of 2009-03-04 (top left) with different learning methods using local information.

6. REFERENCES

[1] 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.

[2] 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.

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

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] V.N. Vapnik and A.Y. 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.

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

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

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.