EoCoE

Energy oriented Center of Excellence

for computing applications

D2.8 - M36

Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression

D2.8 - M36 Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression

E∘C∘E

Project and Deliverable Information Sheet

EoCoE	Project Ref:	EINFRA-676629
	Project Title:	Energy oriented Centre of Excellence
	Project Web Site:	http://www.eocoe.eu
	Deliverable ID:	D2.8 - M36
	Lead Beneficiary:	CEA
	Contact:	Edouard Audit
	Contact's e-mail:	edouard.audit@cea.fr
	Deliverable Nature:	Report
	Dissemination Level:	PU*
	Contractual Date of Delivery:	30/09/2019
	Actual Date of Delivery:	30/09/2019
	EC Project Officer:	Carlos Morais-Pires

* - The dissemination level are indicated as follows: PU – Public, CO – Confidential, only for members of the consortium (including the Commission Services) CL – Classified, as referred to in Commission Decision 2991/844/EC.

Document Control Sheet

Document	Title :	Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression
	ID :	D2.8 - M36
	Available at:	http://www.eocoe.eu
	Software tool:	LaTeX
Authorship	Written by:	Filipa Varino
	Contributors:	Michael Zamo, Liliane Bel, Olivier Mestre, Joel Stein
	Reviewed by:	Edouard Audit, Nathalie Girard - PEC members

2

Contents

List of Figures

List of Tables

Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression

MICHAËL ZAMO

Météo-France, Toulouse, France

LILIANE BEL

AgroParisTech, Paris, France

OLIVIER MESTRE[a] AND JOËL STEIN

Météo-France, Toulouse, France

ABSTRACT

Numerical weather forecast errors are routinely corrected through statistical postprocessing by several national weather services. These statistical postprocessing methods build a regression function called model output statistics (MOS) between observations and forecasts that is based on an archive of past forecasts and associated observations. Because of limited spatial coverage of most near-surface parameter measurements, MOS have been historically produced only at meteorological station locations. Nevertheless, forecasters and forecast users increasingly ask for improved gridded forecasts. The present work aims at building improved hourly wind speed forecasts over the grid of a numerical weather prediction model. First, a new observational analysis, which performs better in terms of statistical scores than those operationally used at Météo-France, is described as gridded pseudo-observations. This analysis, which is obtained by using an interpolation strategy that was selected among other alternative strategies after an intercomparison study conducted internally at Météo-France, is very parsimonious since it requires only two additive components, and it requires little computational resources. Then, several scalar regression methods are built and compared, using the new analysis as the observation. The most efficient MOS is based on random forests trained on blocks of nearby grid points. This method greatly improves forecasts compared with raw output of numerical weather prediction models. Furthermore, building each random forest on blocks and limiting those forests to shallow trees does not impair performance compared with unpruned and pointwise random forests. This alleviates the storage burden of the objects and speeds up operations.

1. Introduction

Numerical weather prediction (NWP) models, although essential for forecasting the dynamics of the atmosphere, are not perfect and may be consistently biased. This is particularly true near the surface (Haiden et al. 2015)

⌀ Denotes Open Access content.

---

[a] Current affiliation: CNRM, UMR 3589, Météo-France/CNRS, Toulouse, France.

Corresponding author address: Michaël Zamo, Direction des Opérations, Météo-France, 42 Ave. Gaspard Coriolis, 31057 Toulouse CEDEX 07, France.
E-mail: michael.zamo@meteo.fr

because processes such as stress and surface heating are not well modeled and because model topography may not be accurate. Furthermore, sources of errors, such as initial condition errors, model errors, and parameterization errors, accumulate in a very intricate way (e.g., initial condition errors are a mixture of model and assimilation errors). These errors may not be easily or quickly corrected through improvement in the knowledge of the atmospheric behavior or in the performance of computers or computing science. A cheap, quick, and efficient means of correcting systematic errors is the so-called model output statistics (MOS; Glahn and Lowry 1972) method, which is used by many national weather services (Wilson

D2.8 - M36 Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression

EoCoE

and Vallée 2002; Baars and Mass 2005; Schmeits et al. 2005; ECMWF 2006; Kang et al. 2011; Zamo et al. 2014). MOS is a statistical postprocessing technique consisting of building a statistical regression function between a predictand or response (what is to be predicted) and predictors or explanatory variables (what is used to make the prediction). Predictors are usually outputs of some NWP model, thus the term MOS. The chosen statistical regression function is then applied to future forecasts to improve their performance in terms of objective scores, such as the root-mean-square error (RMSE) or the mean error.

The predictand in MOS is usually a variable measured at meteorological stations. As a consequence, MOS is mainly applied to station locations, and its performance is evaluated against measurements at those stations. However, forecast users need improved forecasts at arbitrary locations where measurements are not always available. For a national weather service, a most interesting goal is to have MOS available over the grid of some NWP model. To achieve this goal, two possible strategies are 1) to build MOS at station locations and then to grid them, as the National Oceanographic and Atmospheric Administration (NOAA) does (Glahn et al. 2009; Gilbert et al. 2009), or 2) to grid measurements and then to build MOS using this gridded field as the predictand. In this study the second strategy is preferred and described.

Specifically, the aim is to build gridded MOS fields over France for hourly 10-m wind speed forecasts. Wind forecast fields have been selected because of their importance in warning systems and the potential damage that can result (damage to building roofs, fallen tower cranes, and injuries or deaths caused by fallen objects are just some examples). Furthermore, as a result of local phenomena (e.g., slope wind, tunneling), surface wind speed is not the easiest field to interpolate or improve and, as such, it is a good candidate for testing the efficiency of the MOS methods. The same methodology will be applied to other fields, such as wind gusts and temperature. The first step is to build a new wind speed analysis and to demonstrate that it performs better in terms of statistical scores than those operational analyses at Météo-France. The necessity for using a different wind analysis comes from the insufficient availability of operational analyses (every 3 h until April 2015), the requirement to have at least 3 yr of hourly gridded wind speed to train MOS methods, and the opportunity to generate more accurate analyses. The interpolation strategy described in this study has been selected among 48 strategies after an intercomparison led at Météo-France (not shown here). The 48 interpolation strategies varied in their interpolation functions, the information used, and their modeling of the spatial dependence. The second step is to build the best MOS using the new analysis. For that aim, two regression methods are

compared. Both are trained by pooling together the data at nearby grid points (or ''blocks'') and deriving the most parsimonious regression functions while keeping the same forecast performance. By reducing the number of regression functions, this so-called block MOS method is useful in speeding up operations when using MOS over a whole country such as France.

The manuscript is organized as follows. In section 2 the more efficient gridded analysis is introduced and compared against the analysis operational at Météo-France. Section 3 is devoted to building gridded MOS of wind forecasts using the more efficient analysis strategy as the observations. Section 4 sums up our results.

2. Gridding 10-m wind speed measurements

To get gridded fields of 10-m wind speed measurements even where actual measurements are not available, several interpolation strategies exist. The most straightforward approach is to use as the predictand an existing analysis of some NWP model. However, classical variational data assimilation schemes such as 3DVAR (Courtier et al. 1991) or 4DVAR (Courtier et al. 1994) assimilate station measurements. Therefore, an objective verification of such an analysis versus those measurements is not straightforward and may lead to overconfidence in the forecasts' performance, as will be shown later. Furthermore, since assimilation schemes mix in some way forecasts and observations, the obtained analysis could be affected by the forecast bias. As presented in Schaefer and Doswell (1979), it is also possible to use the two-dimensional wind field, interpolating divergence and vorticity instead of the wind vector itself. This may allow imposing physical constraints, such as mass conservation, and using the wind vector instead of the wind speed only. But while it works across a limited domain, this solution requires boundary conditions that may not be trivial. A third efficient method for interpolating measurements is to run a very high-resolution model and find a statistical relationship between measurements and short lead-time forecasts at the same (or nearby) locations. This interpolation function is built for locations where the predictand and the predictors are available and applied to points where only the predictors are known, as presented in Burlando et al. (2013). This approach typically uses an NWP model with a resolution on the order of a few tens of meters. This is not feasible for a whole country as wide as France, but a good compromise could involve using a model over the entirety of France with a grid size of a few kilometers. This statistical interpolation is the approach chosen here, and the results are compared to an analysis existing at Météo-France, which is a kilometer-scale analysis based on 4DVAR assimilation. The methodology is presented in more detail hereafter.

5

a. Methodology

Let us suppose we have at our disposal past predictand and predictor values, at time $t = 1, \ldots, T$ for N^s stations located at sites s_i, where $i = 1, \ldots, N^s$. Let us note that $\mathscr{S}$ is a fine (model) grid covering the region of interest, and $\mathscr{T}$ is a fine temporal grid covering $(1; T)$. Then, for a generic spatiotemporal point (s, t), with $s \in \mathscr{S}$ and $t \in \mathscr{T}$, let us note that $y(s, t)$ and $\mathbf{x}(s, t)$ are the values of the predictand and the vector of predictors, respectively.

Interpolating the predictand consists of building some function f such that $y(s, t) = f[\mathbf{x}(s, t)] + \varepsilon(s, t)$, with ε an interpolation error. The function f is built to have the best generalization capability, that is the lowest possible errors ε over the sites in $\mathscr{S}$. It is fitted locally; that is, for a given spatiotemporal point (s_i, t), the training set $\mathscr{D}(s_i, t)$ is made of a subset of $\{s_1, \ldots, s_{N^s}\} \times \mathscr{T}$ depending on (s_i, t).

Many interpolation strategies can be tried by varying the training set, the family of functions to which f belongs, the choice of the predictors $\mathbf{x}$, and the optional modeling of the error ε. The error can be supposed to be deterministic (Hengl 2007) with no modeling at all. Alternatively, the error can be simulated with statistical models either without spatiotemporal dependence (Hastie et al. 2009; Kuhn and Johnson 2013) or with spatial dependence treated explicitly (Hengl 2007; Cressie and Wikle 2011).

b. Data description

The predictand is the hourly 10-m wind speed defined as the average of the instantaneous wind speed measurements taken during the 10 min before each hour. These measurements are available at 436 meteorological stations over mainland France (named above s_i, with $i = 1, \ldots, N^s$), which are managed by Météo-France. To balance the quantity and quality of the measurements, the retained data are actually measured at heights between 8 and 13 m for stations of environmental class lower than or equal to 3 according to the World Meteorological Organization's Guide to Meteorological Instruments and Methods of Observation (WMO 2008, chapter 1, annex 1.B). For wind speed measurements, environmental class 3 requires that "the mast should be located at a distance of at least 5 times the height of surrounding obstacles" and that "sensors should be situated at a minimum distance of 10 times the width of narrow obstacles (mast, thin tree) higher than 8 m." The mean distance between pairs of nearest stations is 21 km. The study period goes from January 2011 to March 2015.

For the best interpolation strategy described hereafter, the vector of predictors $\mathbf{x}$ at a site s is composed of the position of the site and the most recent wind speed forecast from an NWP model.

- The position of each site $s \in \mathscr{S}$ is specified by its horizontal coordinates (s_x and s_y) in the extended Lambert-93 georeferencing system and its altitude s_z. The value of s_z is obtained by considering the altitude of the nearest point in the BD Alti (http://professionnels. ign.fr/bdalti) digital elevation model (DEM), which is made available through the French geographical institute [Institut national de l'information géographique et forestière (IGN)]. The freely available version of this DEM, which is used in this study, has a resolution of 75 m and covers France only.

- The most recent wind speed forecast from an NWP model is made with Applications de la Recherche à l'Opérationnel à Méso-Echelle (AROME), Météo-France's high resolution NWP model. It is a limited-area, nonhydrostatic model. During the study period, it had a 2.5-km grid size over France (Seity et al. 2011). For one specific site, date, and time, the wind speed forecast comes from the most recent run, excluding the analysis, and it is denoted $W_{\mathrm{AROME}}(s, t)$. Since AROME runs four times per day, the lead times used ranged from 1 to 6 h. As an example, for an interpolation at 1600 UTC, the predictors come from the 1200 UTC run with a lead time of 4 h. The wind speed forecast used at station locations is AROME's forecast bilinearly interpolated from AROME's grid toward these locations.

c. Verification strategy

Since no wind speed measurements are available at grid points, assessment of the interpolation strategy is achieved through cross-validated interpolation toward some test stations. Cross validation consists of splitting the available archive into two subsets: one training set is used to fit the interpolation functions and the test set is used to assess the interpolation performance.

Since cross validation is time consuming, a subset of 150 test stations were chosen, which represented the French topography and hourly wind speed climatology. Ten lists of 15 stations were built as test sets, so that each list gathers stations far enough from one another to ensure that the results are close to those of leave-one-out cross validation. The closest test stations in each list are separated by at least 80 km. Interpolation is done toward each of these 10 test lists separately, and their performance is assessed. Consequently, the training is always done with 421 stations (up to missing data).

Comparing this new analysis against the existing AROME analysis provides an assessment of its usefulness for operational purposes. However, the AROME

6

assimilation scheme already assimilates station measurements, which biases its scores toward better performance. Thus, in order to get an accurate assessment of the analysis performance as an interpolator, 10 AROME assimilations were rerun without assimilating one test set of 15 stations each. Since this reanalysis takes time, it was only run for 120 dates between July 2013 and July 2014, at 1500 UTC, which corresponds to the maximum of the diurnal cycle of wind speed. This reanalysis is referred hereinafter as AROME$_{cv}$, since it is computed with cross validation.

Finally, until April 2015, the AROME analysis was available only every 3 h, whereas MOS is required at an hourly rate. Consequently, a simple reference hourly interpolation method is built by bilinear interpolation of AROME's most recent wind speed forecast, with a lead time of 1–6 h. At some site s with geographical coordinates (s_x, s_y), bilinear interpolation takes as an interpolation function $f[\mathbf{x}(s, t)] = a + bs_x + cs_y + ds_x s_y$. The parameters a, b, c, and d are fitted onto the four nearest AROME grid points from the interpolation point s. If this bilinear interpolation performs better, the retained analysis is simply the most recent wind speed AROME forecast.

For each of these analyses, the interpolation performance is assessed by pooling together the interpolated values in the 150 test stations at the 120 test dates. Classical performance measures are used, such as

- bias,

$$\text{BIAS} = \overline{-\varepsilon(s,t)};$$

- root-mean-square error,

$$\text{RMSE} = \sqrt{\overline{\varepsilon^2(s,t)}}; \quad \text{and}$$

- mean absolute error,

$$\text{MAE} = \overline{|\varepsilon(s,t)|},$$

where $\varepsilon(s, t)$ is the aforementioned interpolation error, and the overbar signifies the mean over all test stations and test dates.

Since RMSE and MAE values alone do not give information about the distribution of errors, specifically about large errors, measures of error dispersion are also computed:

- percentage of absolute errors lower than or equal to w, with $w = 1$ or $4 \, \text{m s}^{-1}$, denoted $\%_{\leq 1}$ and $\%_{\leq 4}$, respectively, and
- quantile of order τ of absolute errors with $\tau = 0.5$ (median) or 0.9, denoted $Q(0.5)$ and $Q(0.9)$, respectively.

d. Results about the best interpolation strategy

The best interpolation strategy among the 48 interpolation strategies previously tested is presented.

First, the training set $\mathscr{D}(s, t)$ is global and run for a fixed time. This means that whatever the interpolation point (s, t) is, the training domain pools all the stations over France but it takes into account only the measurements at time t.

Second, the interpolation function is a mixture of two thin plate regression splines (TPRSs; Wood 2003). This is a special class of generalized additive models (GAMs; Wood 2006). In GAMs the actual predictand is some link function g of the expectation of y, taken as the sum of p functions: $g[\mathbb{E}(y \mid \mathbf{x})] = \sum_{j=1}^{p} f_j(x_j)$, with x_j being one or several components of the predictors vector. Here, the link function is the identity, and the functions f_i are two TPRSs. Indeed, our best interpolation function is simply $f\{\mathbf{x}[s, t; \mathscr{D}(s, t)]\} = \text{tps}[W_{\text{AROME}}(s, t)] + \text{tps}'(s_x, s_y, s_z)$, where tps and tps$'$ are two TPRSs, whose parameters are fitted for each date and time in an automatic way by means of the function gam in the R package mgcv (R Core Team 2015).

Third, the spatial dependence between the errors is not explicitly modeled in this strategy. It appears to be unnecessary since using the AROME wind speed forecast implicitly imposes some structure onto the interpolated field.

Unless otherwise stated, the following results are computed for the 150 test stations, the 120 test dates, and at 1500 UTC.

1) COMPARISON TO REFERENCE AND CROSS-VALIDATED AROME ANALYSES

The two first columns in Table 1 present the measures of performance for the TPRS analysis and the reference. For the whole sample, both analyses are unbiased. However, TPRS performs better than bilinear interpolation for the other measures of performance. The RMSE is improved by 16%, and most of the errors are less than $4 \, \text{m s}^{-1}$ in absolute value.

Table 1 also shows the same measures of performance but for classes defined by the terciles of the wind speed distribution over France during the study period: weak (below $2.9 \, \text{m s}^{-1}$), average (between 2.9 and $4.8 \, \text{m s}^{-1}$), and strong (above $4.8 \, \text{m s}^{-1}$). For the lowest measured wind speeds, TPRS and the reference tend to yield slightly too strong winds (positive bias) and the converse for the strongest measured wind speeds (negative bias). However, the bias remains low. Whatever the wind speed regime, TPRS outperforms bilinear interpolation whatever other performance measure is considered.

Figures 1 and 2 show the evolution of RMSE and BIAS over the time of day for TPRS and reference analyses, computed over the 150 test stations and all of the

D2.8 - M36 Improved Gridded Wind Speed Forecasts by Statistical Postprocessing of Numerical Models with Block Regression

EoCoE

TABLE 1. Measures of performance for TPRS, bilinear reference interpolation (ref.), operational AROME analysis (AROME), and AROME reanalysis computed with cross validation (AROME_{cv}). These results concern 150 test stations and 120 dates at 1500 UTC, for all wind speed values and three different intervals of wind speed measurements. Values in boldface indicate the best performance among TPRS, reference, and AROME_{cv}.

	TPRS	Ref.	AROME	AROME_{cv}
	All wind speed values			
BIAS	**0.0**	0.3	−0.1	**0.0**
MAE	**1.0**	1.2	0.6	1.1
RMSE	**1.4**	1.6	0.8	1.5
$Q(0.5)$	**0.8**	0.9	0.4	**0.8**
$Q(0.9)$	**2.1**	2.5	1.2	2.3
$\%_{\leq 1}$	**63.1**	53.8	86.3	58.3
$\%_{\leq 4}$	**98.8**	97.6	99.6	98.3
	Weak wind (below 2.9 m s^{-1})			
BIAS	0.7	0.8	0.1	**0.5**
MAE	**0.9**	1.1	0.5	**0.9**
RMSE	**1.2**	1.5	0.7	1.3
$Q(0.5)$	**0.7**	0.9	0.3	**0.7**
$Q(0.9)$	**2.0**	2.4	1.0	**2.0**
$\%_{\leq 1}$	**66.9**	56.2	90.1	64.9
$\%_{\leq 4}$	**99.4**	97.7	99.8	98.9
	Medium wind (between 2.9 and 4.8 m s^{-1})			
BIAS	0.1	0.4	−0.1	**0.0**
MAE	**0.8**	1.1	0.5	0.9
RMSE	**1.1**	1.4	0.7	1.2
$Q(0.5)$	**0.7**	0.8	0.3	**0.7**
$Q(0.9)$	**1.7**	2.2	1.0	2.0
$\%_{\leq 1}$	**70.6**	56.8	89.3	63.1
$\%_{\leq 4}$	**99.8**	99.1	100.0	99.5
	Strong wind (above 4.8 m s^{-1})			
BIAS	−0.7	**-0.3**	−0.4	−0.7
MAE	**1.3**	**1.3**	0.7	1.4
RMSE	**1.7**	1.8	1.1	1.8
$Q(0.5)$	**1.0**	**1.0**	0.5	1.1
$Q(0.9)$	**2.7**	2.8	1.6	2.9
$\%_{\leq 1}$	**51.9**	48.8	79.5	46.8
$\%_{\leq 4}$	**97.4**	96.1	99.0	96.5

dates in the study period. The curves may show abrupt changes every 6 h, when the predictors are taken from a different run. This is due to the better performance of the underlying forecast thanks to the proximity of AROME assimilation. Regardless, TPRS performs consistently better than the reference, and its performance shows less variability. This is also true for other performance measures (not shown here).

Table 1 also shows the performance measures of the operational AROME analysis with all stations assimilated and of the AROME_{cv} reanalysis. As an example of the usefulness of this cross-validated reanalysis for assessing the performance of the TPRS analysis, let us note that without blacklisting some stations the operational AROME analysis gets an RMSE of about 0.8 m s^{-1} over the test stations, a significantly better score compared with the actual cross-validated RMSE of 1.5 m s^{-1} (87.5%

higher). This shows the strong local impact of the observations in the assimilation fields.

As for the new analysis, it appears that TPRS actually performs better than the AROME_{cv} reanalysis, whatever the interval of measured wind speeds and the performance measure. Moreover, TPRS is computed very quickly: the complete hourly interpolated wind speed grid from January 2011 to March 2015 required only 4 days of computation at the resolution of AROME (2.5 km) on a standard workstation. This may allow a real-time computation of wind speed analysis to be used routinely and builds a long enough archive to train the MOS methods.

2) SPATIAL STRUCTURES OF GRIDDED MEASUREMENTS

The performance measures used quantify the quality of interpolation strategies but say nothing about the likeliness of the structures represented in the gridded wind field. Figure 3 allows a subjective evaluation of these structures. Figure 3 presents the storm Joachim that hit western Europe in December 2011.

First, TPRS may increase or decrease the wind speed compared with the AROME forecast. As an example, in Fig. 3 the gridded wind speeds with TPRS are lower than those in the forecasts across southwestern France but more variable and stronger in the Pyrenees. The wind speed in the new analysis is also increased at the tip of Brittany and decreased over a large area to the east and southeast of Brittany. This high-impact event has been subjectively evaluated by meteorologists thanks to Météo-France's internal reports of this event. The structures in TPRS have been judged to be more in agreement with reality.

Moreover, the gridded wind speeds, although usually smoother than AROME because of the use of smoothing functions such as TPRS, still exhibit realistic physical structures. This may not be systematic for every interpolation strategy. Indeed, as an example, ordinary kriging led to unrealistic smooth wind speed fields (not shown). In Fig. 3, the wind speed field is more variable throughout the Pyrenees for TPRS than for the AROME forecast. Because AROME only includes 2.5-km-resolution topography whereas the new analysis includes the 75-m-resolution BDAlti topography, this increased spatial variability of the gridded wind speed over the mountains seems to be a positive feature.

Similar results hold for other dates and hours that have been subjectively appraised by Météo-France (not shown).

3) WHY A GLOBAL TRAINING DOMAIN?

A local training domain, containing only of stations within a certain radius around each site s, was used as a sensitivity experiment. This training radius was varied between 20 and 2000 km. Indeed, a variographic study
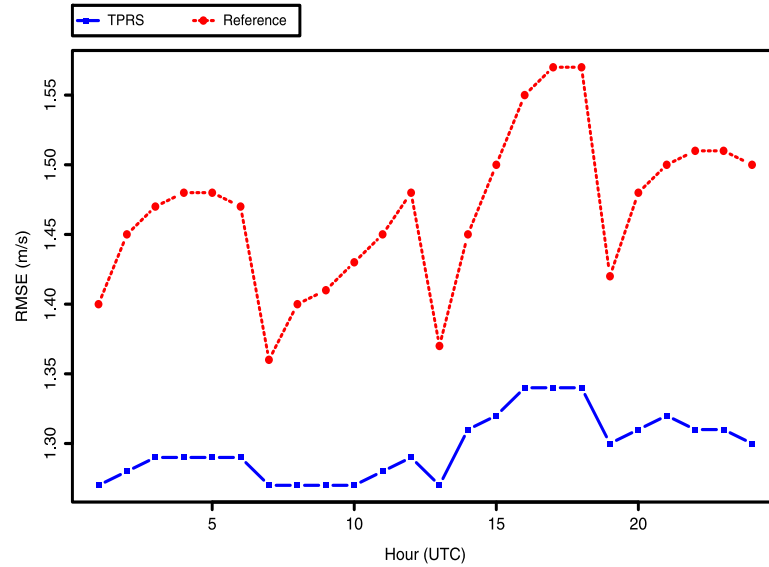
FIG. 1. Evolution of RMSE over time for TPRS and reference interpolation strategies, computed over the 150 test stations.

(not shown here) showed that the correlation length of wind speed measurements is about 50 km, albeit with large differences according to the kind of meteorological and geographical zone (warm sector, tail end of a low, neighborhood of a front, mountains, sloping areas, etc.). It could be expected that, with smaller training domains, the wind speed measurements would be more correlated and the performances improved. But it turns out that these local training domains delivered poorer performances than did a global training domain. It happens that the smaller the training domain, the less numerous the data and the less precise the estimation of the interpolation function, thus the worse interpolation performance (not shown). Inversely, by taking a global training domain, the interpolation method takes the best of all available data at one specific time. To improve performance by reducing the size of the training domain would require a much denser measurement network. In hilly or mountainous
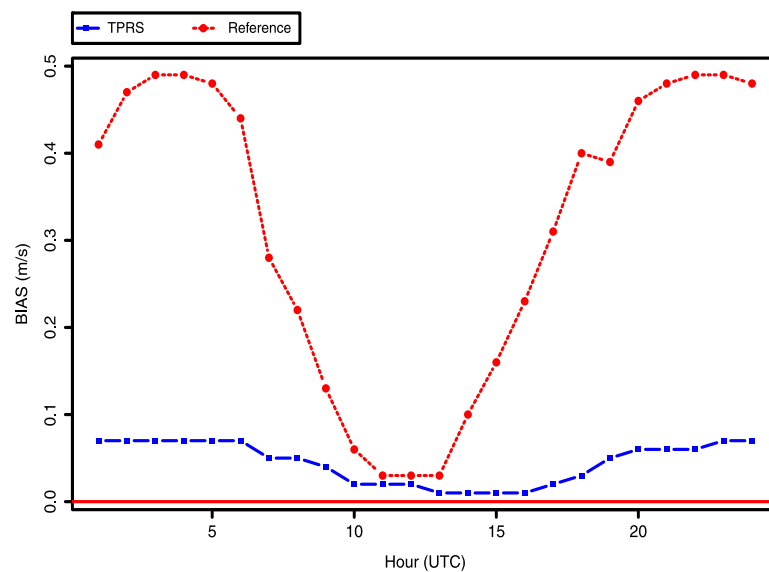

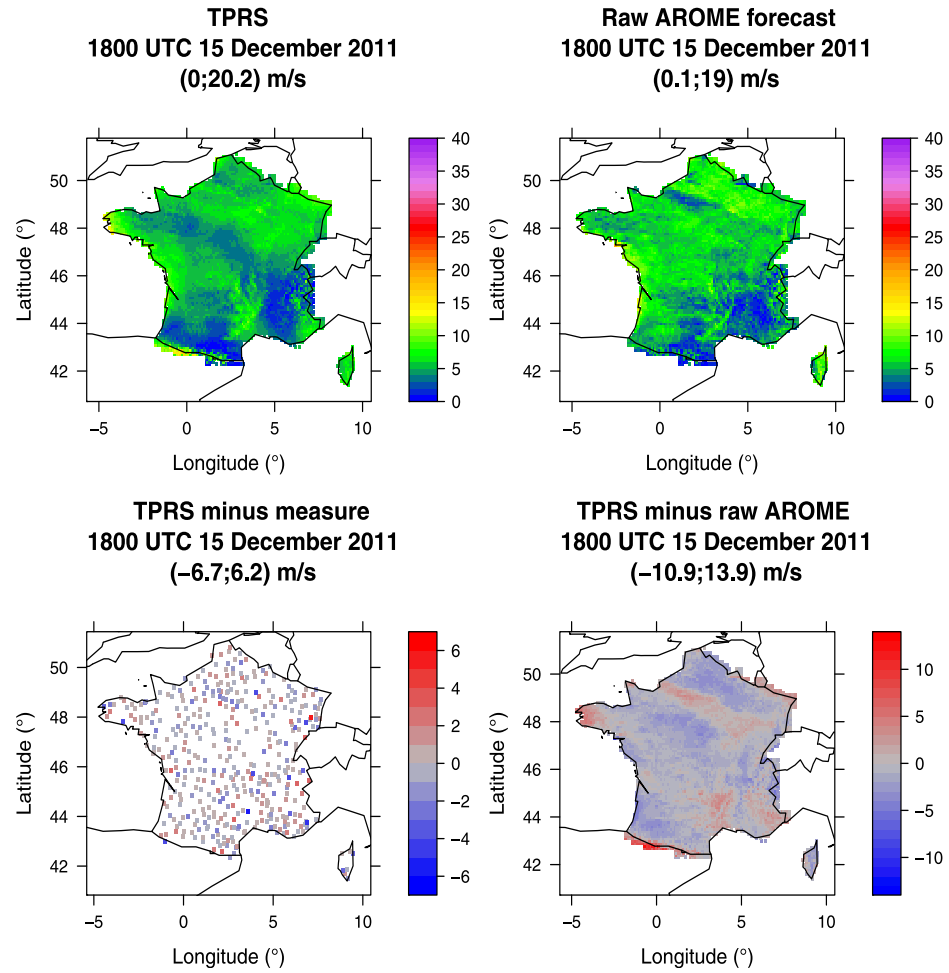FIG. 2. As in Fig. 1, but for the bias.

9

Fig. 3. Results of interpolation at 1800 UTC 15 Dec 2011. (top left) Map of gridded wind speeds with TPRS. (top right) Map of AROME wind speed forecast (run at 1200 UTC, lead time +6 h). (bottom left) Residuals of interpolation at station locations. (bottom right) Differences between TPRS and the AROME forecast. Under each map title is the interval of the corresponding quantity.

areas, with very local topographic effects, this requirement would become unrealistic.

4) FURTHER POSTPROCESSING OF TPRS INTERPOLATION

By construction, TPRS linearly extrapolates as soon as there is a predictor exceeding the values in the training dataset. Because of this linear extrapolation, interpolated wind speeds may reach unrealistic values. Contrary to other interpolation strategies tried, TPRS nearly never exhibits such excessive wind speeds. To filter out and prevent these rare occurrences, a postprocessing of the gridded wind speed fields illustrated before is applied (see example in Fig. 4). The meteorological spline in TPRS, $tps[W_{AROME}(s, t)]$, is constrained at each grid point to be less than $tps[\max_{training}(W_{AROME})]$,

where $\max_{training}(W_{AROME})$ is the maximum AROME forecast in the training dataset. Since this filtering rarely changes the gridded measurements, performance measures of TPRS are not modified.

Finally, a visual comparison of measured values and TPRS interpolation at the station locations showed that for 12 stations, although the new analysis performs better than the AROME analysis, very high errors (up to 80% below the measured value) remain. These stations are situated in hilly areas and exhibit very high values. These features make it very unlikely to develop a good interpolation at these locations. To keep high wind speeds in the new analysis, measurements at all stations are simply copied out to the nearest grid point.

To conclude this section, TPRS is a quick and more efficient alternative to the usual data assimilation scheme
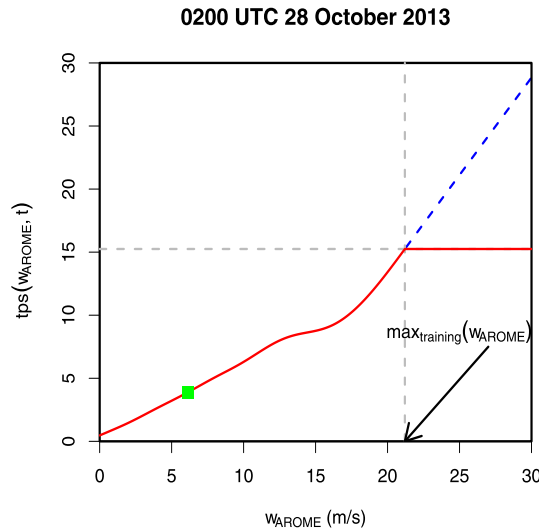
0200 UTC 28 October 2013

FIG. 4. Postprocessing of gridded wind speed to prevent excessive extrapolation in TPRS. The AROME wind speed forecast is W_{AROME}. The blue dashed line is the original meteorological spline component at the chosen grid point. The red continuous line is the postprocessed meteorological spline. The green square is the actual gridded wind speed at the chosen grid point (unchanged in this case).

when creating a long archive of hourly gridded wind speed measurements. TPRS also requires less computational resources and can be run on a standard workstation.

The following section describes how to improve wind speed NWP forecasts using MOS with the new analysis based on TPRS as the predictand (or the response).

3. Improving wind speed forecasts on a grid by block regression

MOS aims at correcting forecasts by means of a regression function r between the variable Y to be predicted and some explanatory variable(s) (or predictors) X that may be NWP model output(s) or any other source of information. This regression function is estimated on an archive of past forecasts and associated observations, and it is then applied to future forecasts to increase their accuracy. It is quite similar to what has been done in the previous section when building an interpolation function, provided that the regression function is applied at future times ($t > T$) instead at nonmonitored locations ($s \notin \{s_i, i = 1, \ldots, N^s\}$).

In this study, two classical regression methods, namely linear modeling and random forests, are compared. The functional kernel regression (Ferraty and Vieu 2006; Ferraty et al. 2012) was also tested but results are not presented since this method is largely outperformed by the two classical regression methods (not shown).

a. Data

The explanatory variables X come from the Action de Recherche Petite Échelle Grande Échelle (ARPEGE; Courtier et al. 1991), Météo-France's global NWP model. ARPEGE is a stretched-grid, hydrostatic NWP model, with a horizontal grid size of 0.1° (about 10 km) over France. It runs every 6 h with hourly lead times of up to 60 or 102 h depending on the run. Table 2 lists the 24 explanatory variables selected for building regression functions of the analyzed wind speed on forecasts. The variables SLP_Adv, SLP_Trend, tpwHPA850, and tpwHPA850_Adv have been chosen as proxies for the synoptic dynamics of the atmosphere. The variables capeins, tH_PCs, ffH_PCs, and tpwHPA850_HVar aim at quantifying the instability of the boundary layer.

The response Y is ARPEGE's wind speed forecast errors relative to the new postprocessed TPRS wind speed analyses, which are presented in section 2d(4). Several attempts showed that the performances were slightly improved when predicting the wind speed error instead of the wind speed itself. Performances are computed for the

TABLE 2. List of ARPEGE's explanatory variables, available for wind speed regression.

Abbrev	Description
ffH10, ddH10f	10-m wind speed and discretized direction (north, south, east, and west)
lat, lon, elevation	Latitude, longitude, and elevation
month	Month as a qualitative variable with 12 categories
capeins	Convective available potential energy
nc, nt, nb, nm, nh	Nebulosities (c, convective; t, total; b, low-altitude clouds; m, medium-altitude clouds; h, high-altitude clouds)
SLP_Adv, SLP_Trend	Advection and 3-h trend of sea level pressure
tpwHPA850, tpwHPA850_Adv, tpwHPA850_Hvar	Potential wet-bulb air temperature at 850 hPa, and its advection (Adv) and horizontal variance (HVar)
tH_PCi, $i = 1, \ldots, 3$	First three components of a principal component analysis of temperature vertical profile (up to 1500 m)
ffH_PC, $i = 1, \ldots, 3$	First three components of a principal component analysis of wind speed vertical profile (up to 1500 m)

11

corrected wind speed forecasts. The regression function is built only on ARPEGE grid points and not on all AROME grid points because of computation time constraints for operational purposes and because of ARPEGE's larger lead time range. Since the new analysis is available on the AROME 2.5-km grid, block MOS for AROME at its full resolution is planned for future applications and are likely to accelerate operations considerably. The study period covers 3 yr, from 1 September 2011 to 31 August 2014.

As a result of long computation times, regression methods have been trained only over 10 spatial domains noted D01–D10 (see Fig. 5). Each domain contains a grid of 9×9 ARPEGE grid points (about $90 \times 90 \, \text{km}^2$). These domains have been chosen so that they represent a large range of conditions of winds and topography over France. Domains D06, D09, D07, and D08 cover increasingly rugged topography. Domains D01, D02, and D03 can be subject to strong local winds, namely marin and cers for the first two domains, and mistral for the third one.

Each regression method is trained separately for lead times of 3, 15, and 48 h for ARPEGE run at 0000 UTC. The lead times have been chosen to cover short and long lead times and, for 15 h, which is representative of the hours of the day that usually have the strongest winds.

b. Block MOS

The following regression methods (Hastie et al. 2009; Kuhn and Johnson 2013) are tested.

- Linear model (Azaïs and Bardet 2006; Weisberg and Fox 2010): the regression function is a second-order polynomial relationship of the explanatory variables $\hat{r}(X) = \beta_0 + \boldsymbol{\beta} \cdot \mathbf{X}^{1,2}$, where β_0 is a real, $\boldsymbol{\beta}$ is a vector of reals, and $\mathbf{X}^{1,2}$ is the vector containing every possible combination of product of explanatory variables of order 1 and 2 (called interactions). The parameters β_0 and $\boldsymbol{\beta}$ are fitted onto the training dataset with an ascending selection of predictors based on the Bayesian information criterion (BIC; Schwarz 1978; Lebarbier and Mary-Huard 2006).
- Random forest (Breiman 2001): this is an average of several regression trees (Breiman et al. 1984). For a single regression tree, the regression function is built through an iterative splitting of available training data into two subsets. Splitting is done according to some threshold of a quantitative explanatory variable or some subset of modalities of a qualitative explanatory variable. The best split is chosen so that the two subsets of response values are the most homogeneous inside each subset and the most dissimilar between one another. The (dis)similarity criterion is the intra- or
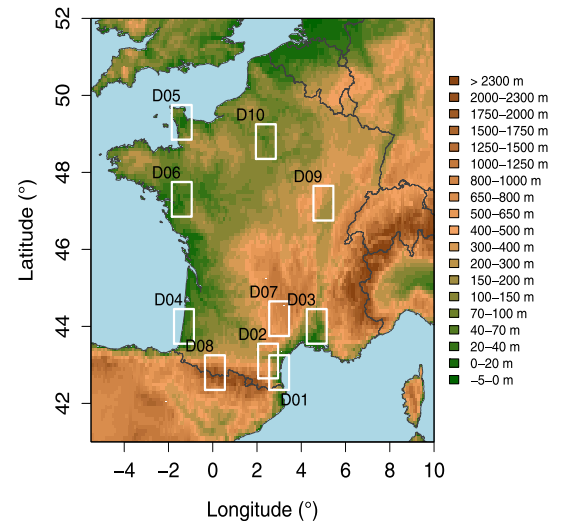

FIG. 5. The 10 training domains used in this study.

between-group variance. Splitting is stopped for some criterion, such as a maximum number of groups, called leaves. The predicted value is then the average of the response values in the leaf. A regression tree usually has a low bias but strongly depends on the training data.

In a random forest, each tree is similar to a regression tree but with two further randomizations. The first randomization is to start each tree from a bootstrapped sample of the training data (Diaconis and Efron 1983). Then each split, or node, of each tree is built from a random subset of the available explanatory variables. The final predicted value is the average of all leaves reached by the value of the vector of explanatory variables. This double randomization makes the trees of the forest more independent and thus decreases the variance of the errors without increasing the bias of each tree. The regression function $\hat{r}$ is an average of blockwise constant functions over the space of the explanatory variables. In this study, the fitted parameters are the number of trees in the forest and the number of predictors tried at each node.

In statistics, more data sometimes imply better inference. Therefore, with the aim of further improving MOS performance, block regression is used. This means that inside each domain, the regression methods described above are trained by pooling data across several grid points. These pooled grid points are collectively called a block. Consequently, one regression function $\hat{r}$ is built for a block and applied to all the grid points inside the block. However, the position of each grid point is available as a predictor through its latitude, longitude, and elevation. If these predictors are selected during the

12

training step, the regression function may actually depend on the gridpoint location. Another advantage expected from block regression is in having fewer models, which may speed up operations.

The size of the block is varied to assess its impact on the MOS performance. The sizes are 1×1 (or pointwise training), 3×3, 5×5, 7×7, and 9×9 grid points. The blocks of sizes 3×3, 5×5, and 7×7 contain the central square with, respectively, 9, 25, and 49 grid points of the domain. If MOS performs better for a nonpointwise block, it is planned to map France with contiguous blocks for using MOS in operations.

To assess forecast performances, the same measures of performance as in section 2c are computed. To compare across the same dataset the training with different block sizes, performance measures are computed for the central 3×3 grid points in each domain. The so-called skill scores are also used: if S_A, S_B, and S_∞ are the measures of performance for forecasts A, B, and a perfect forecast, the associated skill score (SS) is $SS_{A/B} = [(S_A - S_B)/(S_\infty - S_B)] \in (-\infty; 1]$. For RMSE, MAE, $Q(0.5)$, and $Q(0.9)$, $S_\infty = 0$, whereas for $\%_{\leq 1}$ and $\%_{\leq 4}$, $S_\infty = 100$. A positive skill score implies the forecast A yields better performance than forecast B.

Furthermore, the levels of variability among the performances are assessed thanks to threefold cross validation: two years' worth of data serve as the training dataset, with the remaining year being used as a test sample. All three possible combinations of two training years/one test year are tried.

c. Results

1) BEST BLOCK MOS

Figure 6 presents the RMSE of raw ARPEGE forecasts and MOS forecasts built with the two regression methods and different combinations of parameters. The scores are computed for the three test years and with training domains of 1×1, 3×3, 5×5, 7×7, and 9×9 grid points. Figure 6 shows domain D03 with a lead time of 15 h. Whatever the chosen training settings, both MOS methods improve performance over raw ARPEGE forecasts.

Performance levels of random forests are sensitive to the number of trees and number of tried predictors at each node. For a given number of trees and tried predictors, performance is slightly decreased by increasing the block size, but this effect is marginal. The best tuning is therefore to take about six to eight tried predictors and at least 50 trees trained in 3×3 blocks. This optimal setting remains true for other domains and lead times (not shown here). To speed up operations, shallower trees may be used if this does not reduce the forecast

performance. With default settings, the complete random forests have 2300 leaves for each tree, for a 3×3 gridpoint training domain. Constraining trees to a maximum number of leaves has been tested. The best performance is achieved by random forests even with no more than 200 leaves, as shown in Fig. 7 for domain D03 and a lead time of 15 h. However, for some other less common domains and lead times, the minimum optimal number of leaves may be around 500 (not shown). To summarize, the best random forest MOS is obtained by building 200 trees with eight tried predictors at each node and 500 leaves.

As for the linear regression MOS specifically, Fig. 6 shows that its performance varies quite a bit with the training block size. However, the best performance is achieved with pointwise training no matter the domain, lead time, or performance measure (not shown).

In Fig. 6, the best linear model (trained pointwise) and the best random forests apparently deliver similar levels of performance. By showing skill scores for random forests (model A) versus the pointwise trained linear MOS (model B), Fig. 8 shows that forecast performance is improved by several percent with random forests. The only exception is for the percentage of absolute errors lower than $4\,\mathrm{m\,s^{-1}}$ ($\%_{\leq 4}$), where the performance may be diminished when using a random forest compared with using linear regression. Figure 8 also confirms that random forests trained on a 3×3 block yield similar performance compared with pointwise random forests. Thus, even though training random forests on blocks does not improve the forecast performance as could be hoped, it also does not decrease the performance. Skill scores computed for other domains and/or lead times confirm that a random forest is usually a better choice than linear regression by a few percent, except for $\%_{\leq 4}$, where the best MOS is not always the same (see, e.g., Fig. 8). In conclusion, the best MOS method is to use a random forest with from six to eight tried predictors, 200 trees, 500 leaves, and a 3×3 block training.

On a more qualitative note, MOS successfully corrects the tendency of the raw model to overestimate wind speed, as illustrated in Fig. 9 with a smoothed scatterplot. As can be seen, the MOS scatterplot is much more concentrated along the first bisecting line than the raw forecast. This line corresponds to the point set of perfect forecasts. This improvement is obvious whatever the strength of the gridded wind speed. These results hold for every other domain or lead time (not shown).

2) PERFORMANCE AT STATION LOCATIONS

Table 3 shows performance measures of forecasts bilinearly interpolated at the locations of the meteorological stations inside the 10 training domains. Scores are
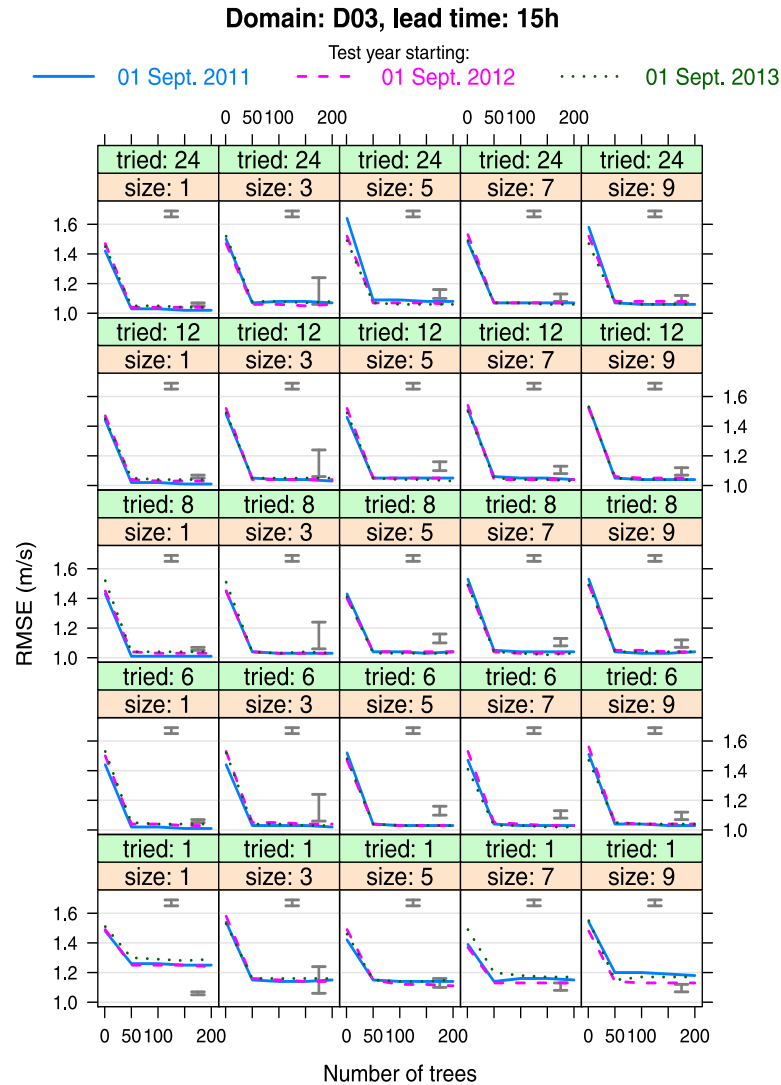
13

FIG. 6. RMSEs for several MOS methods and settings, along with raw ARPEGE forecasts. In each panel, lines show the evolution of the performance of the random forest with the number of trees, for a specific training block size and number of tried predictors at each node and for the three test years. In each panel, vertical bars indicate the interval of variation over the three test years of ARPEGE performance (left bars) and block MOS with a linear regression (right bar). For a linear model and a random forest, the training domain can be of sizes 1×1, 3×3, 5×5, 7×7, and 9×9 grid points from left to right. Scores are computed over one year of testing (starting on 1 Sep), over the 3×3 central points of domain D03 and for 15-h lead time.

computed relative to measurements at those stations by pooling forecasts for all three test years and for all three lead times. Only five stations were included in any of the 10 training domains. For the raw AROME forecasts, a 48-h lead time is actually a 6-h lead time for the 1800 UTC run, since AROME does not yield forecasts beyond 36-h lead time. However, the valid dates are the same for MOS at 48-h lead times and this lagged raw AROME.

The scores show that a random forest delivers better overall performance than do interpolated raw ARPEGE forecasts, for a training domain of pointwise or 3×3 blocks. Furthermore, random forests yield similar or better levels of performance than do interpolated forecasts from Météo-France's high-resolution model AROME. Concerning the bias, whereas random forests have a negative bias and AROME is unbiased, the bias of the
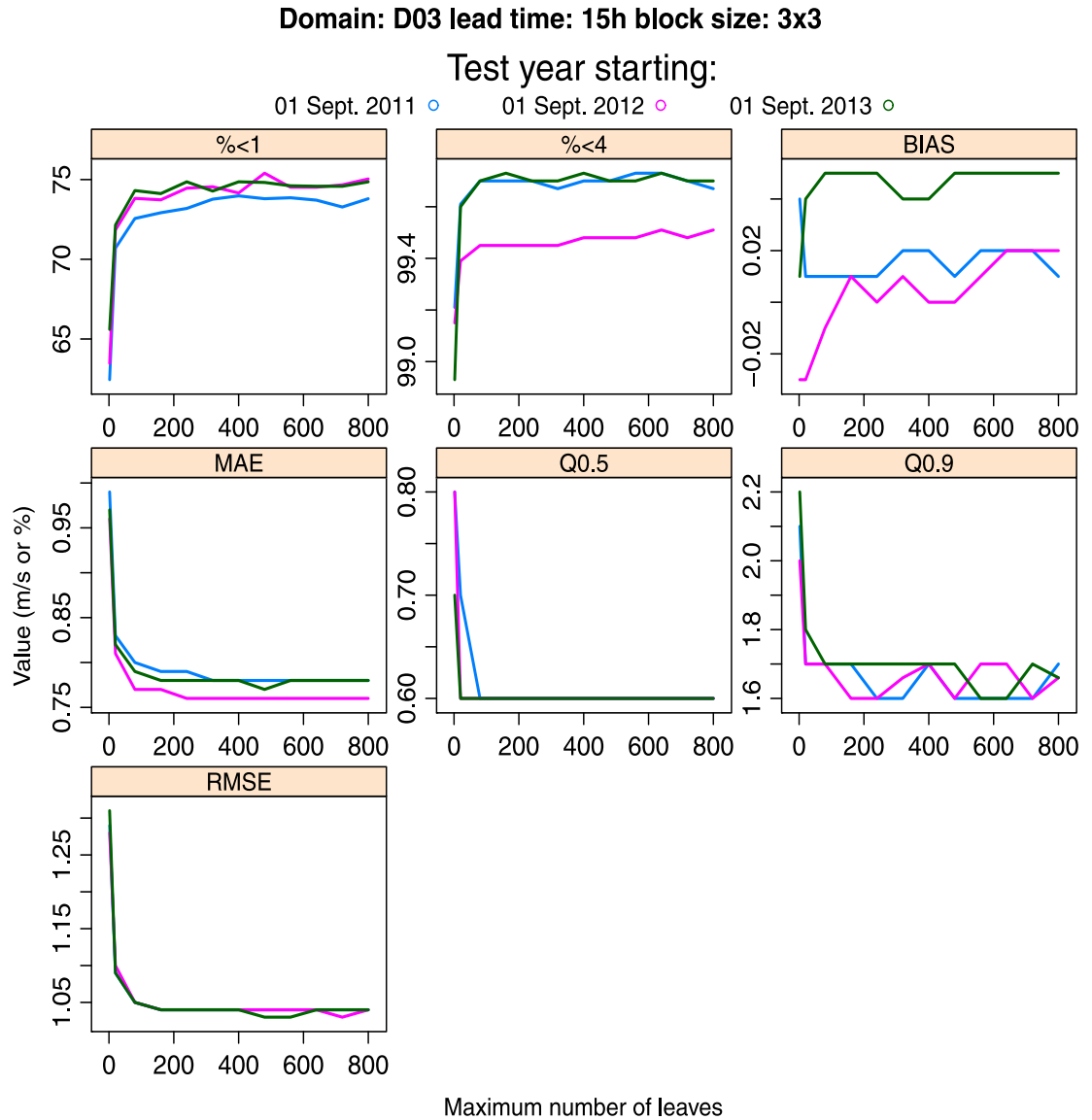
14

Domain: D03 lead time: 15h block size: 3x3

Fig. 7. Variation of measures of performance for random forests, with varying numbers of maximum allowed nodes. Random forests are built with 200 trees and eight tried predictors at each node. Shown is domain D03, at 3-h lead time, and with a 3×3 training block.

random forests remains low (only $-0.3\,\mathrm{m\,s^{-1}}$). For 48-h lead times, MOS is as good as raw AROME at 6-h lead time, an improvement of 42 h.

However, the results vary at the scale of single stations. Table 4 shows the scores obtained for a station picked at random for different lead times. For this station situated in domain D03, random forests achieve much better performance than ARPEGE or AROME for a lead time of 3 h. At a lead time of 15 h, spatially interpolated random forests still get the upper hand over

raw AROME but the differences are slightly reduced. At a lead time of 48 h (6 h for raw AROME), random forests and AROME yield similar results. Over the five stations in the training domains, the results are variable, even though the random forests usually deliver results that are at least as good as those of the interpolated AROME forecasts. Regardless, ARPEGE never prevails. Since the sample is small (only five stations), further investigation would be necessary to assess the best choice among the interpolated
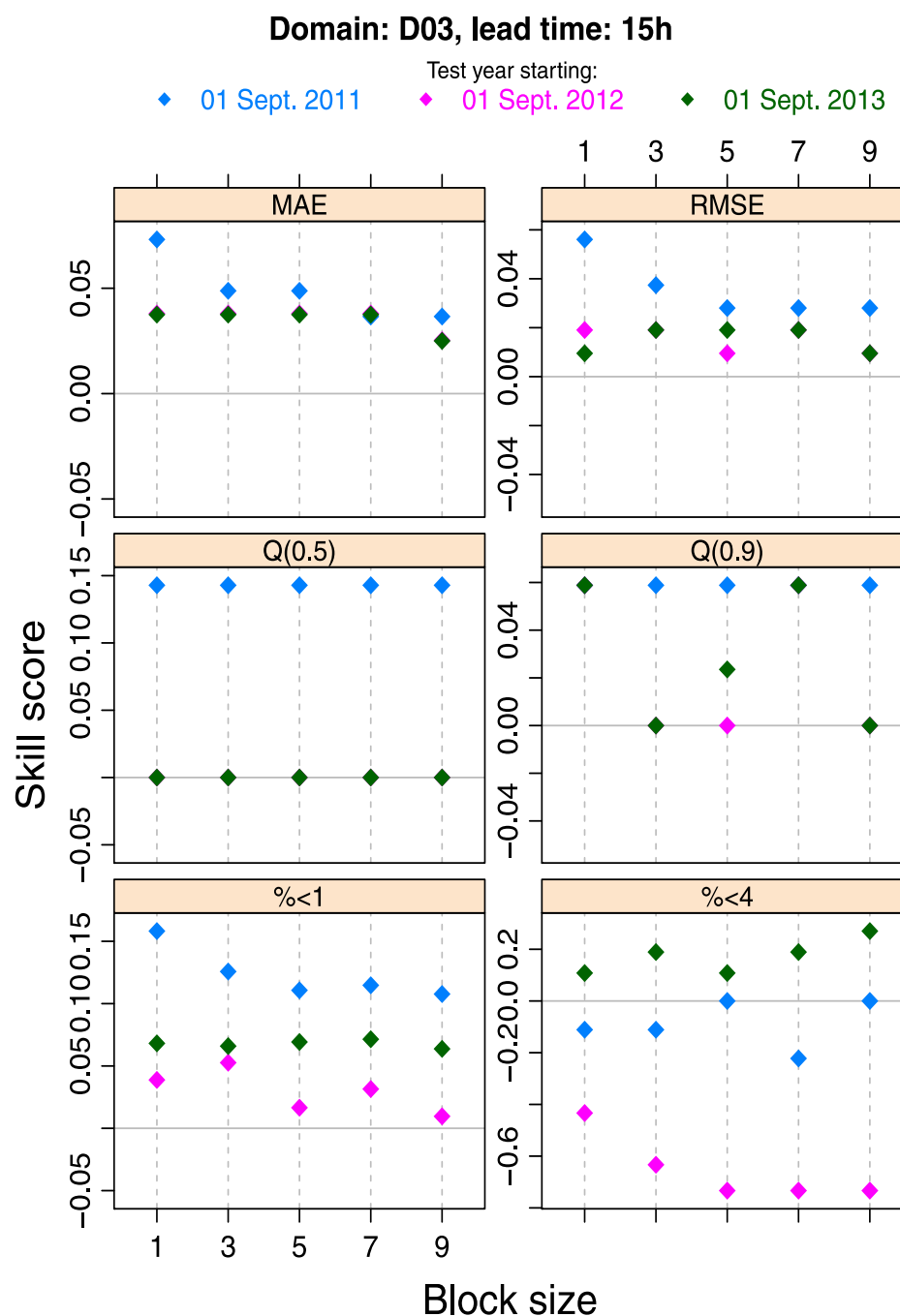
15

FIG. 8. Evolution of skill scores with the size of the training block for random forest MOS, with pointwise trained linear MOS as a reference. Two hundred trees and eight tried predictors are used to train the random forest with several sizes of training block.

forecasts. This will first require us to build MOS forecasts for all of the chosen grids over the whole of France. This will be done for future applications at Météo-France, but such a training scheme will require weeks. Nevertheless, those first results point to interpolating MOS forecasts trained on a 3×3 blocks as a good solution for getting improved forecasts at station locations.
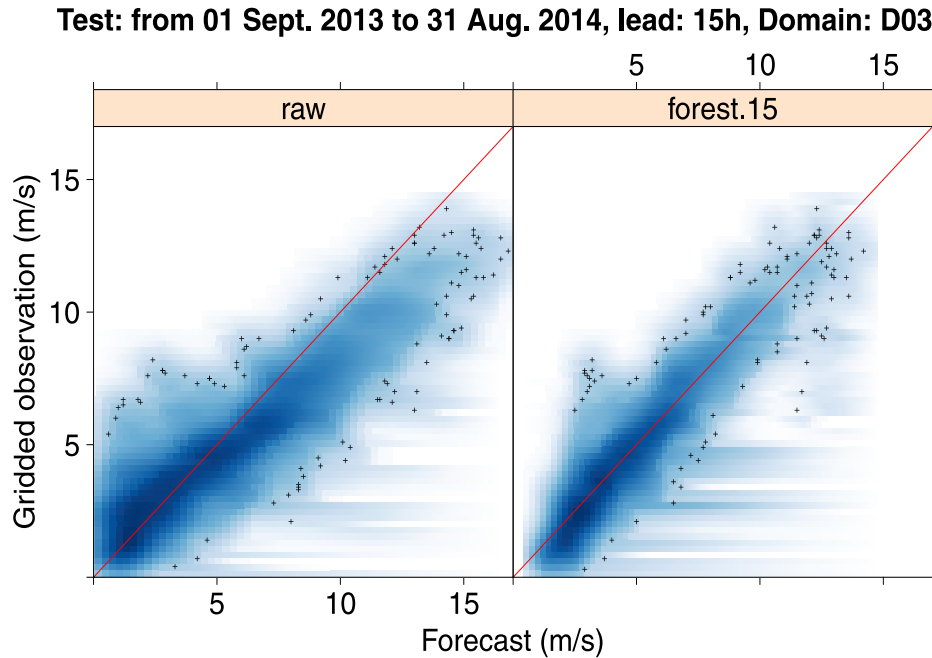
FIG. 9. Smoothed scatterplots of gridded observations against (left) raw forecasts and (right) random-forest-based MOS forecasts. The darker the blue, the denser the points. The red oblique line is the first bisecting line. These scatterplots are for test grid points in domain D03, for the test year starting on 1 Sep 2013, at 15-h lead time, and training on 3×3 grid points. The random forest is built with 200 trees and eight input variables randomly drawn at each node.

3) SPEEDING UP OPERATIONS

Running MOS on a grid with thousands of grid points may be time consuming, at the training stage and during operations. One purpose of block regression is to build fewer regression models to accelerate memory loading during operations. Indeed, since the prediction with random forests is very quick, a limiting factor for operational purposes is the loading time of the models in memory. For the linear models, only the regression coefficients β have to be saved on disk, with a disk occupation of a few kilobytes. A random forest object can be much bigger if it is not optimized. In our case a random forest trained on one grid point amounts to 2 MB (for a total of 18 MB for a nine-gridpoint domain), whereas a random forest trained on a 3×3 domain requires 12 MB, one-third less. Additionally, a shallow forest with 200 trees, eight tried predictors, and only 500 leaves, trained over 3×3 grid points, requires only about 5 MB for each domain, a further reduction of 60%. Removing the components of random forest objects, as stored by the R statistical software, that are deemed to be unnecessary for prediction leads to a final storage size of 1.7 MB on disk.

To compare loading times for several MOS models as stored in R, the above objects have been loaded from disk 300 times for the 10 studied domains. Figure 10 shows that the linear model objects load much more quickly

(about 15 ms for the 10 domains) than pointwise trained random forest objects (a few seconds cumulated over the 10 domains). However, combining block regression, shallow trees, and the removal of unnecessary components allows dividing the loading times of random forests by a factor 10. Since the complete mapping of France requires about 830 domains, the loading time would be about half a minute for the whole country. This still causes the random forest longer to load than linear models, but it

TABLE 3. Measures of overall performance of bilinearly interpolated forecasts at station locations. The forecasts are MOS based on the random forest method, with a training domain of size 1×1 or 3×3, raw ARPEGE forecasts, and raw AROME forecasts. Scores are computed by pooling together forecasts over the three test years, every station inside any study domain, and the three lead times (3, 15, and 48 h). For AROME, 48-h lead time is actually 6-h lead time, since AROME forecasts do not extend up to 48 h. Boldface values indicate the best performance.

| | Random forest | | | |
	1×1	3×3	ARPEGE	AROME
BIAS	−0.3	−0.3	0.2	**0.0**
MAE	**1.2**	1.3	1.7	1.4
RMSE	**1.7**	1.8	2.3	1.9
$Q(0.5)$	**0.9**	1.0	1.2	1.0
$Q(0.9)$	**2.7**	2.8	3.9	3.1
$\%_{\leq 1}$	**54.0**	52.8	45.1	50.8
$\%_{\leq 4}$	**96.5**	96.1	90.7	94.9

17

TABLE 4. As in Table 3, but for one station in domain D03 and for each lead time.

| | Random forest | | | |
	1×1	3×3	ARPEGE	AROME
	3-h lead time			
BIAS	**0.2**	**0.2**	1.5	0.9
MAE	**1.0**	**1.0**	1.7	1.3
RMSE	**1.3**	**1.3**	2.1	1.6
$Q(0.5)$	**0.9**	**0.9**	1.5	1.0
$Q(0.9)$	**2.0**	**2.0**	3.4	2.6
$\%_{\leq 1}$	60.1	**60.4**	33.8	48.2
$\%_{\leq 4}$	**99.5**	**99.5**	95.3	99.0
	15-h lead time			
BIAS	**-0.1**	0.2	0.3	**-0.1**
MAE	**1.2**	**1.2**	1.4	**1.2**
RMSE	**1.5**	**1.5**	1.8	1.6
$Q(0.5)$	**1.0**	**1.0**	1.1	**1.0**
$Q(0.9)$	**2.3**	**2.3**	2.9	2.5
$\%_{\leq 1}$	51.5	**52.2**	45.1	49.3
$\%_{\leq 4}$	98.8	**99.1**	96.3	97.4
	48-h lead time			(6 h)
BIAS	**0.2**	**0.2**	1.4	0.8
MAE	**1.2**	**1.2**	1.8	**1.2**
RMSE	**1.5**	**1.5**	2.2	**1.5**
$Q(0.5)$	**1.0**	**1.0**	1.5	1.1
$Q(0.9)$	**2.4**	**2.4**	3.6	2.5
$\%_{\leq 1}$	49.8	**50.1**	35.5	47.6
$\%_{\leq 4}$	98.5	98.5	92.7	**99.0**

is compatible with operational constraints. As seen above, this acceleration is achieved without reducing the overall performance of the forecast.

4. Conclusions

Accurate wind speed forecasts are crucial for decision-making in weather-related activities or for weather warnings by national and regional weather services. NWP models provide forecasts that are not exempt from errors. Since these errors are not completely random, statistical postprocessing methods, known as MOS, can be used to improve future forecasts by using regression functions fitted onto past forecasts and associated observations. To apply those methods to wind speed forecasts at gridpoint locations, a new gridded analysis of wind speed measured at meteorological stations is built. An internal comparison of 48 interpolation strategies led at Météo-France showed the best hourly analysis is based on thin plate regression splines. This regression is very parsimonious with only two additive components: a first one with the most recent wind speed forecast of the high-resolution model AROME as the only input and the second one with a correction based on the three-dimensional coordinates of the points. By cross validation, it is shown that this new analysis performs consistently better than the available AROME analysis
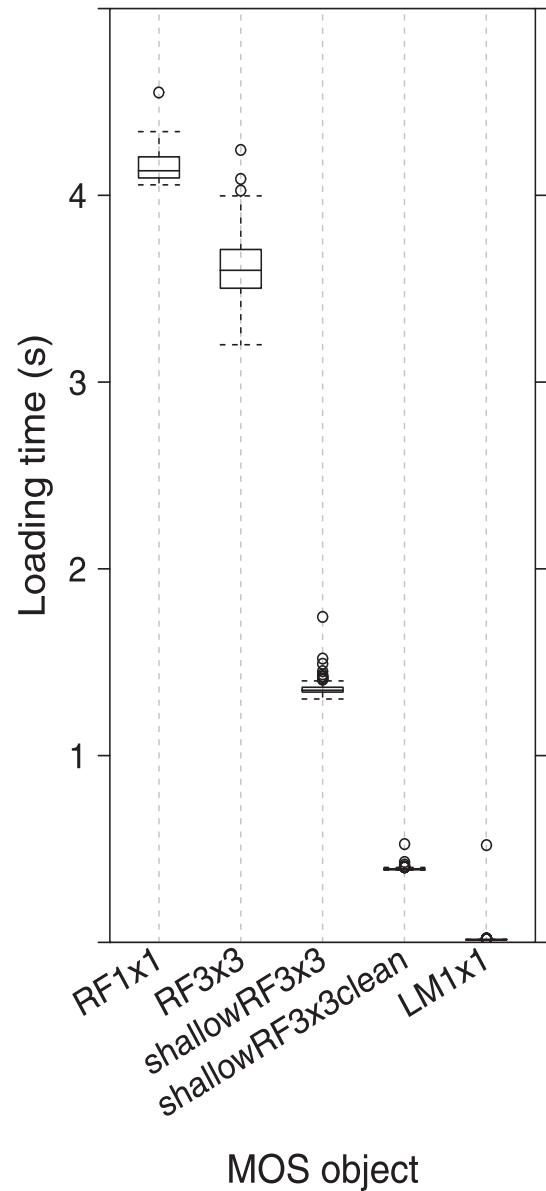

FIG. 10. Boxplots of 300 loading times for the whole set of R objects over the 10 training domains for several MOS models: complete random forest trained pointwise (RF1 $\times$ 1), complete random forest trained over a 3×3 block (RF3 $\times$ 3), shallow random forest trained over a 3×3 block (shallowRF3 $\times$ 3), shallow random forest trained over a 3×3 block and with the removal of unnecessary elements for prediction in objects (shallowRF3 $\times$ 3clean), and a pointwise linear model (LM1 $\times$ 1).

while retaining the realistic structures of the wind speed fields thanks to the use of the AROME forecast in the interpolation function. This allows us to build an archive of gridded wind speeds over France with a 2.5-km grid size starting in January 2011 and ending in March 2015.

18

This new analysis is used to build improved wind speed forecasts of Météo-France's 10-km NWP model, ARPEGE, over France. The use of classical regression methods shows that ARPEGE forecasts are easily and greatly improved by all regression methods. The best MOS is based on random forests. The best combination of parameters for this model is shown to be not very sensitive: taking more than 200 trees and trying from six to eight predictors at each node is sufficient. Furthermore, random forests can be trained by pooling together data from nearby grid points without degrading performances. Also, the trees in the optimal random forests need not be very deep in order to achieve the best performance. These last remarks lead to building less numerous and shallower random forests. After removing unnecessary components in R random forest objects, the storage resources and loading times of the random forests are reduced by a factor of 10. The time to produce MOS forecasts is mainly determined by the loading time of all the random forests into memory. Thanks to their reduced size and number, this operation can be done in a reasonable time period (about half a minute) that enables its application in everyday operations. This MOS method with random forests trained over blocks is currently being made operational at Météo-France by covering France with contiguous blocks. A new analysis of gusts has also been developed, and block MOS for gusts is being made operational as well.

REFERENCES

Azaïs, J., and J. Bardet, 2006: *Le Modèle Linéaire par l'Exemple: Régression, Analyse de la Variance et Plans d'Expériences Illustrés avec R, SAS et Splus.* Dunod, 326 pp.

Baars, J. A., and C. F. Mass, 2005: Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Wea. Forecasting*, **20**, 1034–1047, doi:10.1175/WAF896.1.

Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, doi:10.1023/A:1010933404324.

——, J. Friedman, R. Olshen, and C. Stone, 1984: *Classification and Regression Trees.* Chapman and Hall/CRC, 368 pp.

Burlando, M., P. De Gaetano, M. Pizzo, M. P. Repetto, G. Solari, and M. Tizzi, 2013: Wind climate analysis in complex terrains. *J. Wind Eng. Ind. Aerodyn.*, **123**, 349–362, doi:10.1016/j.jweia.2013.09.016.

Courtier, P., C. Freydier, J. Geleyn, F. Rabier, and M. Rochas, 1991: The ARPEGE project at Météo-France. *Proc. Workshop on Numerical Methods in Atmospheric Models*, Vol. 2, Reading, United Kingdom, ECMWF, 193–231.

——, J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quart. J. Roy. Meteor. Soc.*, **120**, 1367–1387, doi:10.1002/qj.49712051912.

Cressie, N. A. C., and C. K. Wikle, 2011: *Statistics for Spatio-Temporal Data.* Wiley Series in Probability and Statistics, Vol. 465, J. Wiley and Sons, 624 pp.

Diaconis, P., and B. Efron, 1983: Computer intensive methods in statistics. *Sci. Amer.*, **248**, 116–130, doi:10.1038/scientificamerican0583-116.

ECMWF, 2006: Application and verification of ECMWF products in member states and co-operating states. ECMWF, 158 pp. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2006/9218-application-and-verification-ecmwf-products-member-states-and-co-operating-states.pdf.]

Ferraty, F., and P. Vieu, 2006: *Nonparametric Functional Data Analysis: Theory and Practice.* Springer Science and Business Media, 260 pp.

——, I. Van Keilegom, and P. Vieu, 2012: Regression when both response and predictor are functions. *J. Multivariate Anal.*, **109**, 10–28, doi:10.1016/j.jmva.2012.02.008.

Gilbert, K. K., B. Glahn, R. L. Cosgrove, K. L. Sheets, and G. A. Wagner, 2009: Gridded model output statistics: Improving and expanding. Preprints, *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, Omaha, NE, Amer. Meteor. Soc., 6A.6. [Available online at https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154285.htm.]

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529, doi:10.1175/2008WAF2007080.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, doi:10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Haiden, T., and Coauthors, 2015: Evaluation of ECMWF forecasts, including 2014-2015 upgrades. ECMWF Tech. Memo. 765, 51 pp. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2015/15275-evaluation-ecmwf-forecasts-including-2014-2015-upgrades.pdf.]

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The Elements of Statistical Learning.* Springer, 745 pp.

Hengl, T., 2007: A practical guide to geostatistical mapping of environmental variables. 2nd ed. UR 22904 EN Scientific and Technical Research Series Rep., Office for Official Publications of the European Communities, 271 pp.

Kang, J.-H., M.-S. Suh, K.-O. Hong, and C. Kim, 2011: Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pac. J. Atmos. Sci.*, **47**, 199–211, doi:10.1007/s13143-011-0009-8.

Kuhn, M., and K. Johnson, 2013: *Applied Predictive Modeling.* Springer, 600 pp.

Lebarbier, É., and T. Mary-Huard, 2006: Une introduction au critère BIC: Fondements théoriques et interprétation. *J. Soc. Fr. Stat.*, **147**, 39–57.

R Core Team, 2015: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 3452 pp. [Available online at https://www.R-project.org.]

Schaefer, J. T., and C. A. Doswell, 1979: On the interpolation of a vector field. *Mon. Wea. Rev.*, **107**, 458–476, doi:10.1175/1520-0493(1979)107<0458:OTIOAV>2.0.CO;2.

Schmeits, M. J., K. J. Kok, and D. H. Vogelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Wea. Forecasting*, **20**, 134–148, doi:10.1175/WAF840.1.

19

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, doi:10.1214/aos/1176344136.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Mon. Wea. Rev.*, **139**, 976–991, doi:10.1175/2010MWR3425.1.

Weisberg, S., and J. Fox, 2010: *An R Companion to Applied Regression*. Sage Publications, 472 pp.

Wilson, L. J., and M. Vallée, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting*, **17**, 206–222, doi:10.1175/1520-0434(2002)017<0206:TCUMOS>2.0.CO;2.

Wood, S. N., 2003: Thin plate regression splines. *J. Roy. Stat. Soc.*, **65B**, 95–114, doi:10.1111/1467-9868.00374.

——, 2006: *Generalized Additive Models: An Introduction with R*. Texts in Statistical Science, Vol. 66, CRC Press, 410 pp.

WMO, 2008: *Guide to Meteorological Instruments and Methods of Observation*. Commission for Instruments and Methods of Observations Tech. Rep. 8, WMO, 681 pp.

Zamo, M., O. Mestre, P. Arbogast, and O. Pannekoucke, 2014: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Sol. Energy*, **105**, 792–803, doi:10.1016/j.solener.2013.12.006.

20