Hrvatsko meteorološko društvo
Croatian Meteorological Society

HRVATSKI METEOROLOŠKI ČASOPIS
CROATIAN METEOROLOGICAL JOURNAL

Časopis je izdan povodom 20. godišnjice ALADIN projekta u Hrvatskoj

50

Hrv. meteor. časopis	Vol. 50	p. 1-144	ZAGREB	2015

Hrvatski meteorološki časopis – Croatian Meteorological Journal, 50, 2015., 121–136

Original scientific paper
Izvorni znanstveni rad

POST-PROCESSING OF ALADIN WIND SPEED PREDICTIONS WITH AN ANALOG-BASED METHOD

Post-procesne ALADIN prognoze brzine vjetra metodom analoga

IRIS ODAK PLENKOVIĆ[1], LUCA DELLE MONACHE[2], KRISTIAN HORVATH[1], MARIO HRASTINSKI[1], ALICA BAJIĆ[1]

1. Meteorological and Hydrological Service, Grič 3, 10000 Zagreb, Croatia
1. Državni hidrometeorološki zavod, Grič 3, 10000 Zagreb, Hrvatska
2. Research Applications Laboratory, NCAR, Boulder, CO, USA

odak@cirus.dhz.hr

Recieved 27 May 2015, in final form 21 August 2015

Abstract: In this paper, different post-processing methods are described and evaluated for deterministic and probabilistic point-based 10-m wind speed forecast over Croatia. These methods are applied to forecasts of operational high-resolution dynamical adaptation model (DADA) run with 2 km horizontal resolution to address the following question: which point-based post-processing method is the best suited for wind forecasting in the operational suite at DHMZ (Meteorological and Hydrological Service of Croatia).
The verification procedure includes several metrics computed considering wind speed as continuous, categorical and probabilistic predictand. Those metrics were used to optimize the configuration, and to test both the deterministic and probabilistic prediction performance. This study shows that deterministic analog-based predictions (AnEn) improve the correlation between predictions and measurements while reducing forecast error better than using Kalman filter based predictions (KF), even though KF shows better bias reduction. The best results are achieved when forecasting the mean of analog ensemble or the Kalman filter of the mean of analog ensemble. Probabilistic AnEn predictions are properly dispersive, while having better resolution, discrimination and skill than forecast generated via logistic regression. These results encourage the potential use of AnEn in an operational environment at the location of meteorological stations, as well as at wind farms.

Keywords: Short-term wind speed forecasting, Analog forecast, Kalman-filtering, Logistic regression, Verification, Complex terrain

Sažetak: U ovom radu opisano je nekoliko različitih post-procesnih metoda prognoziranja 10-m brzine vjetra koje su potom testirane na nekoliko lokacija u Hrvatskoj. Metode koriste izlaz operativnog visoko-rezolucijskog (2 km) modela dinamičke adaptacije (DADA). Cilj rada je odgovoriti na pitanje: koja je od korištenih post-procesnih metoda najpogodnija za operativno korištenje na Državnom hidrometeorološkom zavodu (DHMZ).
Da bi se optimiziralo i testiralo metode, u verifikacijskom procesu brzina vjetra razmatrana je kao kontinuirani, kategorički i probabilistički prediktand. Pokazano je da deterministička prognoza putem analoga poboljšava korelaciju između prognoze i mjerenja, istovremeno smanjujući pogrešku uspješnije od referentne prognoze temeljene na Kalman filtriranju (KF), iako potonja bolje uklanja pristranost prognoze. Najbolji rezultati se postižu kod prognoziranja srednjaka ansambla analoga i Kalman filtriranog srednjaka. Testirani probabilistički produkt prognoze putem analoga pokazao se prikladno disperzivan, bolje rezolucije, diskriminacije i vještine u odnosu na referentnu prognozu baziranu na logičkoj regresiji. Rezultati sugeriraju mogućnost operativnog korištenja, kako za lokacije mjernih postaja, tako i za lokacije vjetroelektrana.

Ključne riječi: Kratkoročna prognoza brzine vjetra, metoda analoga, Kalman-filtriranje, logička regresija, verifikacija, kompleksni teren

1. INTRODUCTION

Even though the skill of numerical weather predictions has improved at both global and regional scales, they are still affected by imperfect boundary and initial conditions, simplification of physical processes and numerical errors. This is especially the case in operational models that are constrained by the available computing capacity and time necessary to produce forecast. For these reasons, besides improving the model itself (i.e., using higher resolution or better parametrization), it is useful to develop post-processing methods that reduce model errors at locations where measurements exist.

The Analog-based method is a state-of-the-art technique used for point-based forecasting. It is based on finding the most similar past numerical weather predictions (analogs) over several variables (predictors), and then forming an analog ensemble (AnEn) out of the corresponding observations. Besides improving a deterministic forecasting system, there is also need for reliably expressing its uncertainty. The AnEn method can be used to produce probabilistic forecasts also, where the forecast probability density function may be estimated from the members of the AnEn.

Analog-based methods have been explored by a number of studies. Pioneering contribution of Van den Dool (1989) revealed the ability to predict the forecast skill of an NWP model, as indicated by a strong spread-skill relationship in a 10-member AnEn. The author used analyses over a localized area and then used the 12-h subsequent analysis to each analog as a plausible 500 hPa height forecast. Afterwards, various procedures have been formulated, including different predictors and analog selection criteria. Application include idealized cases with low-order models (Ren and Chou 2006), general circulation modeling (Gao et al. 2006, Ren and Chou 2007), long-range weather (Xavier and Goswami 2007), short-term visibility (Esterle 1992), El Niño Southern Oscillation index forecasts (Drosdowski 1994), calibration of probabilistic predictions (Hamil and Whitaker 2006) etc. More recently Klausner et al. (2009) proposed the "similar day method", Panziera et al. (2011) used radar observations for very short-term orographic precipitation predictions, the -nearest neighbors approach was tested in hydrology (Hopson and Webster 2010) and seasonal weather predictions (Wu et al. 2012). However, due to excessive degrees of freedom of the problem at stake, the use of analogs for forecasting of meteorological fields is limited.

Recently, Delle Monache et al. (2011) proposed two analog-based post-processing methods to improve deterministic NWP forecasts of 10-m wind speed based only on the time-series of numerical weather predictions and observations at a single site. They demonstrated that this approach increases correlation and reduces random and systematic errors. The same methodology was used for predicting other variables as well (i.e., Djalalova et al. 2015, showed similar result predicting $PM_{2.5}$ concentration). Delle Monache et al. (2013) also explored benefits from using the AnEn approach to produce probabilistic 10-m wind speed and 2-m temperature forecasts. Probabilistic analog-based prediction was also applied to gain wind resource estimates (Vanvyve et al. 2015), and to predict wind energy (Alessandrini et al. 2015b; Junk et al. 2015) and solar energy (Alessandrini et al. 2015a).

In this study the mean and median of the analog ensemble are tested and compared to a linear, adaptive and recursive Kalman filter post-processing approach (see Delle Monache et al. 2006, 2008, 2011). Two combinations of these post-processing approaches were tested as well. The first one is to apply the Kalman filter algorithm to the time series of deterministic analog-based forecasts, which will produce new deterministic forecast called KFAN. Another way is to apply Kalman filtering to the historical set of (starting) model forecasts in the analog space, ordered from the worst to the best analog. The best analog after the correction will be deterministic forecast named KFAS. Also, as shown by Delle Monache et al. (2013) AnEn can be used to generate a probabilistic prediction from a deterministic forecast. The latter is compared with probabilistic predictions from a logistic regression approach.

In section 2 the post-processed predictions are described, section 3 describes the experiment setup and datasets used and section 4 introduces the verification procedure used.. In sec-

tion 5 the results are presented, followed by the conclusions in section 6.

2. POST-PROCESSING METHODS

2.1. Kalman filter

The Kalman filter (KF) is a recursive algorithm used to estimate a signal from noisy measurements. In this post-processing method the recent past forecasts and observations (past prediction errors) are used by the KF to estimate the future bias in the current raw forecast.

Kalman (1960) showed that the optimal recursive predictor of forecast bias x_t at time t (derived by minimizing the expected mean square error) can be written as a combination of the previous bias estimate and the previous forecast error y_t:

$$\hat{x}_{t+\Delta t|t} = \hat{x}_{t|t-\Delta t} + K_t(y_t - \hat{x}_{t|t-\Delta t}) \quad (1)$$

where the hat (^) indicates the estimate. The weighting factor K_t, called Kalman gain, can be calculated from:

$$K_t = \frac{p_{t-\Delta t} + \sigma^2_{\eta,t}}{(p_{t-\Delta t} + \sigma^2_{\eta,t} + \sigma^2_{\varepsilon,t})} \quad (2)$$

where p is the expected mean-square error, that can be computed as follows:

$$p_t = (p_{t-\Delta t} + \sigma^2_{\eta,t})(1 - K_t) \quad (3)$$

and $\sigma^2_{\eta,t}$, $\sigma^2_{\varepsilon,t}$ are variances of the noise term and the unsystematic error term, respectively. The KF algorithm will quickly converge for any reasonable estimate of p_0 and K_0. Additional details of the procedure and algorithm can be found in Delle Monache et al. (2006). Advantages of KF approach are the short training period and the ability to adapt to changing synoptic conditions. A disadvantage is that it is less likely to predict extreme bias events (Delle Monache 2011)

2.2. Logistic regression

Logistic regression is a model output statistics (MOS) technique specifically designed to produce probabilistic forecasts. It considers a past relationship between predictor variable(s) and the predictand to produce a forecast of the predictand given the predictors' values in the current forecast cycle. The predictand is the probability of a predefined event, such as wind speed greater than 5ms^{-1}. A nonlinear function is fit to past pairs of the predictor(s) and the predictand, that, as an observed value, takes on a probability of either 1.0 (event occurred) or 0.0 (event did not occur). The relationship is linear in terms of logarithm of the odds ratio:

$$\ln\left(\frac{p_i}{1-p_i}\right) = b_0 + b_1 x_1 + \cdots b_k x_k \quad (4)$$

where p_i is the probability of the event i, b_j are regression coefficients, x_j are predictor values and is the number of predictors. So logistic regression can be fitted using ordinary linear regression, except that the predictand is binary (left side is either ln(0) or ln(∞)) and the distribution used is binomial (Wilks 2011).

2.3. Analog ensemble

The AnEn seeks to estimate a separate probability distribution ($f(\cdot)$) of the observed value of the predictand variable given a model prediction, that can be represented as $f(y|x^f)$, where, for a given time and location, y is the observed future value of the predictand variable and $x^f=(x_f^1,x_f^2,...,x_f^k)$ contains the values of k predictors from the deterministic model prediction at the same location and over a time window centered at the same time. This method generates samples of y given x^f via three main steps using historical data, called the analog training period. In that period both the NWP deterministic prediction and the verifying observation are available. Analogs (the best-matching historical forecasts for the current prediction) are sought independently at each location and for each lead time, so an analog may come from any past date within the training period, i.e., a day, week or several months ago. The quality of the analog (i.e. closeness of the match) is determined by the following metric:

$$\|NWP_t\ A_{t'}\| = \sum_{i=1}^{N_A} \frac{w_i}{\sigma_{fi}} \sqrt{\sum_{j=-\bar{t}}^{\bar{t}} \left(F_{i,t+j} - A_{i,t'+j}\right)^2} \quad (5)$$

where F_t is the current NWP deterministic forecast valid at the future time t at a station location, $A_{t'}$ is an analog at the same location and with the same forecast lead time, but valid at a past time t', N_A and w_i are the number of physical variables used in the search for analogs and their weights, respectively. Furthermore, σ_{fi} is the standard deviation of the time series of past forecasts of a given variable at the same location, $\bar{t}$ is equal to half the number of additional times over which the metric is computed, and $F_{i,t+j}$ and $A_{i,t'+j}$ are the values of the forecast and the analog in the time window for a given variable. The analog searching algorithm is highly flexible, and allows the search to occur over a time window of any specified width. The verifying observation for each analog is an actual member of the analog ensemble (AnEn). The assumption is that if analog forecasts are found, their errors will likely be similar to the error of the current forecast and it can be inferred from theirs (Delle Monache et. al. 2011.).

Once AnEn is formed, it can be used to produce a probabilistic forecast (probability of a predefined event) or a deterministic one. Examples of deterministic forecasts are the AnEn mean and median.

It is possible to use a combination of analog forecasts and the Kalman-filter. One way to do it is to apply the KF algorithm to time series of an AnEn based deterministic forecast. Another way is to apply Kalman filtering to the historical set of (starting) model forecasts in the analog space, ordered from the worst to the best analog. In the latter case the correction for the current forecast gives more weight to the analog forecasts closer to it. The goodness of the analog match is defined by the same metrics as previously mentioned (Djalalova et al, 2015).

3. DATASETS AND EXPERIMENT SETUP

The post-processed forecasting methods described in section 2 are tested at 14 locations in Croatia covering different climatological regions (Figure 1). Boxplots of observed data show that the average wind speed value is around 3 ms^{-1} with the maximum value at noon and the minimum at midnight (Figure 2). Post-processing methods are applied to 10-m wind speed forecasts from an operational high-resolution dynamical adaptation model (DADA). Dynamical adaptation procedure (Žagar and Rakovec, 1999) takes the output fields from the operational limited-area mesoscale model ALADIN (Aire Limitee Adaptation Dynamique developement International, ALADIN International Team, 1997) run with 8 km horizontal resolution using hydrostatic dynamics with spectral solver on 37 hybrid sigma-pressure vertical levels (Tudor et al. 2013). Initial conditions computed using variational data assimilation for the upper air fields and optimum interpolation for surface fields (Stanešić 2011), the lateral boundary conditions are taken from the ARPEGE (Action de Recherche Petite Echelle Grande Echelle) global model run operationally in Meteo France. Using high resolution terrain representation, DADA dynamically adapts wind fields to higher horizontal resolution (2 km) grid by integrating the model to reach a quasi-stationary state (Ivatek-Šahdan and Tudor 2004). DADA is run on 15 levels in the vertical. Vertical levels in the planetary boundary layer are at the same heights as in the operational ALADIN model

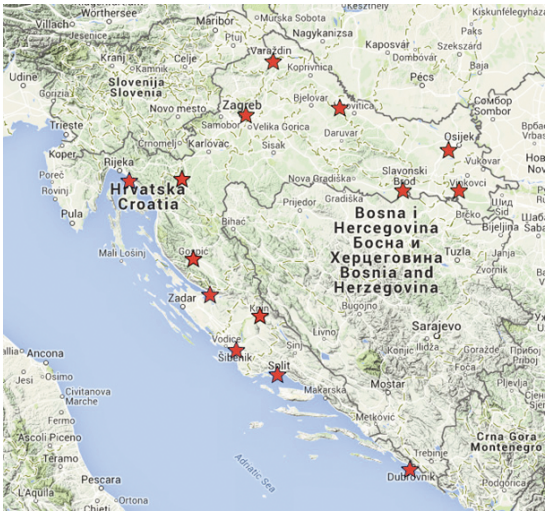

Figure 1. Spatial distribution of the 14 stations providing the observations of 10-m wind speed used in this study.

Slika 1. Prostorna distribucija 14 mjernih postaja s kojih su preuzeti podaci o 10-m brzini vjetra korišteni u ovom radu.

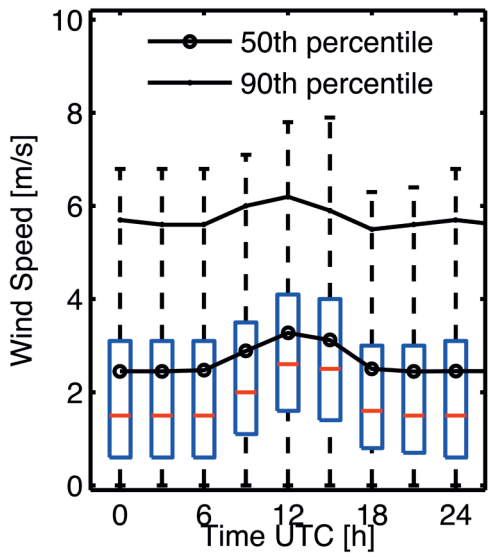

Figure 2. Boxplots of observed data (outliers are not shown) measured at 14 stations in Croatia during the 2010-2012 period, as a function of UTC. Black lines represent 50th and 90th percentile that are used as thresholds in the categories of verification procedure.

Slika 2. Dijagrami s pravokutnikom predstavljaju podatke mjerene na 14 meteoroloških postaja u Hrvatskoj u razdoblju od 2010. do kraja 2012. godine za različito doba dana. Stršeći podaci nisu prikazani, a crnim linijama su istaknuti 50. i 90. percentile koji se u kategorijskoj verifikaciji koriste kao granice kategorija.

with 37 levels (the lowest level is about 17 m above ground), but the ones in the upper troposphere and stratosphere are reduced. The wind field is interpolated to the height of measurements using the stability functions (Geleyn 1988). In dynamical adaptation, turbulence is the only parametrization scheme used, while contributions of the moist and radiation processes are neglected. This cost-effective refinement was run operationally twice a day (00 and 12 UTC run) for 72 h ahead with a 3 hour model output frequency. In complex terrain dynamical adaptation improves numerical near-surface wind predictions, as described in numerous studies such as Tudor and Ivatek-Šahdan (2002), Ivatek-Šahdan and Tudor (2004), Ivatek-Šahdan and Ivančan-Picek (2006), Bajić (2003), Bajić et al. (2007, 2008), Horvath et al. (2009, 2011) etc.

Numerical weather prediction and observation datasets in the period 2010-2012 are divided to training and verification periods. Training period is from 2010 to 2011, while 2012 was used for the verification period. For every location the most representative grid point is chosen from the 4 closest ones to the measuring station and only 00 UTC run is used. Since DADA model does not include moist and radiation physics, only physical variables regarding wind fields are included in the search for the best analogs: wind speed and direction logarithmically interpolated to 10 m height, vorticity and divergence at the lowest vertical level (~17 m). Weight assigned to wind speed and direction is 1 and 0.8 for the other two variables Time frame window used to find the most similar analogs included a 6-hour time window (one time step before/after, so $\bar{t}$ in eq. (5) equals 1. To choose an appropriate number of AnEn members (N), RMSE, RCC and bias results are averaged for all locations and all of the lead times, and their dependency on N is investigated (Figure 3). Mean confidence intervals estimated with bootstrapping are shown for every forecast and every measure. The DADA model, KF, and KFAS do not depend on N. The AnEn mean, AnEn median and KFAN show similar behavior - by increasing N correlation improves, while bias slightly worsens. The RMSE is reduced at first, but then enlarges again for N>15. Because the error and bias growth for large N, the optimal number of AnEn members selected is 15, which is used hereinafter.

Probabilistic forecast predicts a probability of an event occurring. Wind speed exceeding 5 ms^{-1} is chosen as an event, because it is round number with less than 20 % probability to occur climatologically. This means that it is not a common event, while it still occurs often enough not to make some measures sensitive to sample size unstable (i.e. Brier skill score).

4. VERIFICATION PROCEDURE

4.1. Evaluation of deterministic forecast performance

To evaluate the performance of different deterministic post-processing methods wind speed forecast can be considered as continuous or categorical predictand. Considered as a continuous variable, wind speed forecasts are evaluated using bias, root-mean-square-error (RMSE) and Spearman correlation coefficient

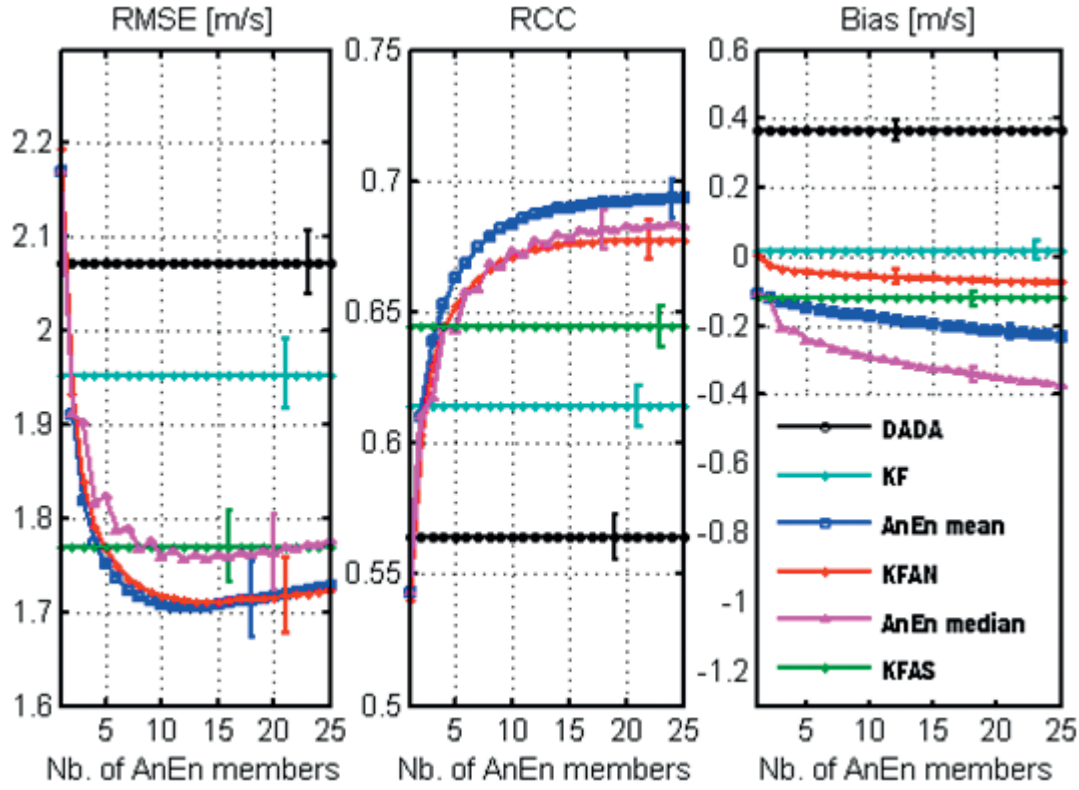

Figure 3. Root-mean-square-error, ranked correlation coefficient and bias dependency on the number of analog ensemble members (N) for the 3 different analog-based techniques averaged over lead times and 14 locations during 2012. Results are compared to 3 deterministic forecasts that do not depend on N. Mean values of the 95% bootstrap confidence intervals are indicated by the error bars.

Slika 3. Korijen srednje kvadratne pogreške, koeficijent korelacije ranga i pristranost u odnosu na broj odabranih članova ansambla (N) za 3 prognoze koje koriste analoge. Rezultati se odnose na 2012. godinu i usrednjeni su za sve postaje i sva nastupna vremena prognoze, te je istaknut prosječni 95 %-tni interval pouzdanosti. Ravne linije odnose se na prognoze koje ne ovise o N.

("rank correlation", RCC). Unlike Pearson correlation coefficient, the RCC is a nonparametric statistic, allowing a nonlinear relationship between predictions and observations. It is a robust and resistant alternative to Pearson correlation, appropriate if dealing with non-Gaussian distributed variables such as wind speed (Wilks 2011; Jolliffe and Stephenson 2011).

Wind speeds are also divided in 3 categories: breeze (or no wind at all), moderate wind and strong wind, depending on climatology. For each lead time thresholds are determined as the 50th and 90th percentile, so they vary due to the diurnal cycle. Categorical verification procedure includes the following metrics:

$$FBias = \frac{F_i}{O_i} \tag{6}$$

$$CSI = \frac{FO_i}{F_i + O_i - FO_i} \tag{7}$$

PCC

where F_i represents the number of events corresponding to category i forecasted, O_i is the number of events corresponding to category i observed, while FO_i is the number of correctly forecasted event corresponding to category i.

Frequency bias (FBias) measures the tendency to forecast too often (FBias greater than 1) or too rarely (FBias less than 1) a particular category (Wilks 2011; Jolliffe and Stephenson 2011).

Critical success index (CSI) measures the fraction of observed forecast events that were correctly predicted. It can be thought of as the accuracy when correct negatives have been removed from consideration, therefore, CSI is only concerned with forecasts that count. Sensitive to hits, it penalizes both misses and false alarms and does not distinguish the source of forecast errors. CSI depends on climatological frequency of events (poorer scores for rarer events) since some hits can occur purely due to random chance. Perfect value is 1 (Wilks 2011; Jolliffe and Stephenson 2011).

Polychoric correlation coefficient (PCC) measures association of forecasts and observations in the contingency table. The bell-shaped bivariate normal density function is assigned to the contingency table and the domain of the bivariate normal density function is cut into rectangles corresponding to the cells of the contingency table. The polychoric correlation coefficient is the parameter value of the bivariate normal density function for which the volumes of the discretized bivariate standard normal distribution is equal to the corresponding joint probabilities of the contingency table. For that purpose values are transformed to standard normal deviates (and thresholds, accordingly). Values for PCC vary between -1 and 1 (1 is for perfect forecast). More details can be found in Juras and Pasarić (2006).

4.2. Evaluation of probabilistic forecast performance

The joint distribution of the forecasts and observations is of fundamental interest with respect to the verification of forecasts. Since it can be difficult to use the joint distribution directly, two factorizations are used. The first one, which is called calibration-refinement factorization, is conditional on the particular value of the forecasts. Attributes related to calibration-refinement factorization use subsets are (Murphy 1993; Wilks 2011):

- Reliability is a measure of the conditional bias of the forecasts or the agreement between forecast probability and mean observed frequency. It is calculated as a weighted average of the squared differences between the forecast probabilities and the relative frequencies of the observed event conditional on forecast probability in each subsample:

$$REL = \frac{1}{n}\sum N_{bin}\left(p(o_1|f_{bin}) - f_{bin}\right)^2. \quad (8)$$

  For perfectly reliable forecasts, the subsample relative frequency is exactly equal to the forecast probability in each subsample and equals zero.

- Resolution is the ability of a forecast to resolve the set of sample events into subsets with characteristically different frequencies. Mathematically, the resolution term is a weighted average of the squared differences between subsamples of relative frequencies of the observed event conditional on forecast probability and the overall sample climatology relative frequency:

$$RES = \frac{1}{n}\sum N_{bin}\left(p(o_1|f_{bin}) - p_{clim}(o_1)\right)^2 \quad (9)$$

  If the forecasts sort the observations into subsamples having substantially different relative frequencies than overall sample climatology, resolution term will be large, which is a desirable situation.

Sharpness is an attribute of the forecasts alone, without regard to their corresponding observations. It characterizes the unconditional distribution of the forecasts f_{bin} in the calibration-refinement factorization. Forecasts that rarely deviate much from the climatological value of the predictand exhibit low sharpness. Thus, sharpness attribute measures tendency to forecast extreme values ("climatology" is not sharp). Sharp forecasts will be accurate only if forecasts also exhibit good reliability (Wilks, 2011).

Uncertainty *(UNC)* depends only on the variability of the observations. It exhibits zero value (minimum) when the climatological probability is either zero or one, and maximum value when the climatological probability is 0.5.
The attributes diagram plots the observed frequency against the forecast probability, where the range of forecast probabilities is divided into K bins (for example, 0-5%, 5-15%, 15-25%, etc.). The diagonal line indicates perfect reliability (average observed frequency equal to predicted probability for each category), and the horizontal line represents the climatological frequency. Sharpness, which is a property of the forecasts only, is diagnosed via a re-

liability diagram by plotting how often probability corresponding to each bin is forecasted (relative frequency).

The reliability is indicated by the proximity of the plotted curve to the diagonal, so the smaller value, the better the forecast is. The deviation from the diagonal gives the conditional bias. If the curve lies below the diagonal line that indicates over-forecasting (probabilities are too high), while points above the line indicate under-forecasting (probabilities are too low). The flatter the curve in the reliability diagram, the less resolution it has. A forecast of climatology does not discriminate at all between events and non-events, thus has no resolution. The reliability diagram is conditioned on the forecasts (i.e., given that X was predicted, what was the outcome?). It is a good partner to the relative operating characteristic, which is conditioned on the observations (Wilks 2011; Jolliffe and Stephenson 2011).

The Brier skill score (BSS) measures the improvement of the probabilistic forecast relative to a reference forecast (climatology), therefore taking uncertainty into account:

$$BSS = (RES - REL)/UNC \quad (10)$$

where 0 indicates no skill when compared to the reference forecast, while 1 indicates perfect skill. This score should always be applied to a sufficiently large sample for which the sample climatology of the event is a representative of the long term climatology. The rarer the event, the larger the number of samples is needed to stabilize the score (Wilks 2011; Jolliffe and Stephenson 2011).

The relative operating characteristic skill score (ROCSS) is calculated using relative operating characteristic (ROC) curve (Figure 4). The ROC is created by plotting the probability of detection as a function of false alarm rate (false alarms / observed no, also known as probability of false detection), using a set of increasing probability thresholds (for example, 0.05, 0.15, 0.25, etc.) to make the yes/no decision. The ROC is conditioned on the observations (i.e., given that Y occurred, what was the corresponding forecast?), unlike attributes diagram that is conditioned on the forecasts. While attributes diagram shows full joint distribution of forecasts and observations for probability forecasts in terms of calibration-refinement factorization, the ROC diagram shows full joint distribution in terms of likelihood-base rate factorization. ROC measures the ability of the forecast to discriminate between two alternative outcomes and a good ROC is indicated by a curve that goes close to the upper left corner (low false alarm rate, high probability of detection). The area under the ROC curve is frequently used as a skill score:

$$ROCSS = \frac{A - A_{RANDOM}}{A_{PERFECT} - A_{RANDOM}} = \frac{A - 1/2}{1 - 1/2} = 2A - 1 \quad (11)$$

where A represents area under the ROC curve, A_{RANDOM} is the area underneath the diagonal (0.5) and $A_{PERFECT}$ is 1. Hence, $ROCSS$ ranges from 0 (forecast with no skill) to 1 (perfect forecast) (Wilks 2011; Jolliffe and Stephenson 2011).

An ensemble is statistically consistent when its members are indistinguishable from the truth. If so, an observation ranked among the corresponding ordered ensemble members is equally likely to take any rank i in the range i=1,2,..., N+1, where N is the number of ensemble members. Collecting the rank of the observation over a number of cases and plotting the results generates a rank histogram, that is flat (i.e., uniform rank probability of

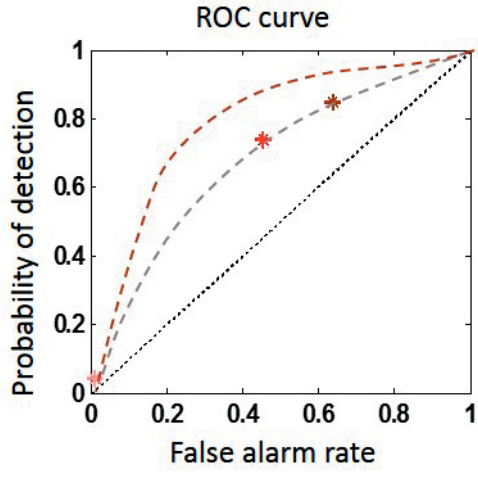

Figure 4. Theoretical example of a ROC curve.

Slika 4. Teoretski primjer ROC krivulje.

[1/(*N*+1)] for a statistically consistent ensemble). Missing rate error (MRE) measures the distance of the first and last (two) bin(s) from the ideal case - so the smaller is the better. More info on MRE can be found in Appendix A of Eckel and Mass (2005).

Statistical consistency over all forecast lead times can be calculated following the general definition that the normalized mean square error of the ensemble mean should match the average ensemble variance. Comparing the square root of those two statistics over all forecast lead times produces a dispersion diagram that shows if an ensemble is properly dispersive. If $\mu_N = \frac{1}{N}\sum_{i=1}^{N} f_i$ is the ensemble mean and $\sigma_N^2 = \frac{1}{N}\sum_{i=1}^{N}(f_i - \mu_N)^2$ is the ensemble variance, then the normalized mean square error of the ensemble mean is computed as:

$$AnEn\ mean\ RMSE = \frac{N}{N+1}\sqrt{\frac{1}{J}\sum_{j=1}^{J}(\mu_j - o_j)^2} \quad (12)$$

and the spread (SPRD) as:

$$SPRD = \sqrt{\frac{1}{J}\sum_{j=1}^{J}\sigma_N^2} \quad (13)$$

where is the number of analogs chosen, while is the number of forecasts produced (Talagrand et al. 1997; Eckel and Mass 2005; Hamill 2001).

5. RESULTS

5.1. Deterministic forecasting

The DADA model overestimates the wind speed minimum at midnight (Figure 5). All post-processing methods shown reduce the

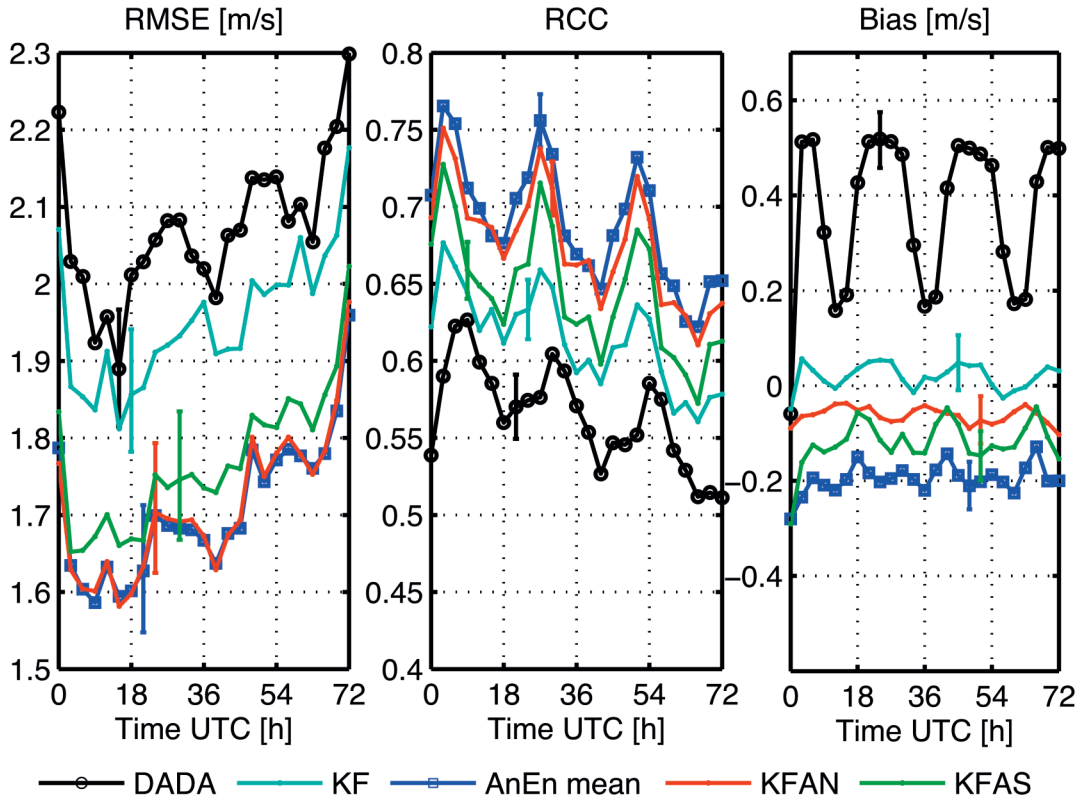

Figure 5. Root-mean-square-error, ranked correlation coefficient and bias dependency on forecast lead time for 5 different deterministic forecasts averaged over 14 locations during 2012. Mean value of the 95% bootstrap confidence intervals are indicated by the error bars.

Slika 5. Korijen srednje kvadratne pogreške, koeficijent korelacije ranga i pristranost u odnosu na nastupno vrijeme prognoze za 5 determinističkih prognoza. Rezultati se odnose na 2012. godinu i usrednjeni su za sve postaje, te je istaknut prosječni 95 %-tni interval pouzdanosti.

bias and have much smaller diurnal variations (results for AnEn median forecasting are similar, but are not shown to avoid clutter). The best results are for the KF based methods (KF and KFAN). Besides the decreasing trend for long lead times, all forecasting systems are less correlated in the afternoon than during the night and in the morning. Although post-processing did not affect that trend, after the correction forecasts are more correlated, especially in case of AnEn mean and KFAN. Similarly to RCC, RMSE is increasing for longer lead times, with a superimposed diurnal error cycle consisting in an increase of errors during the night and decrease in the afternoon for all cases shown. Maintaining the same general

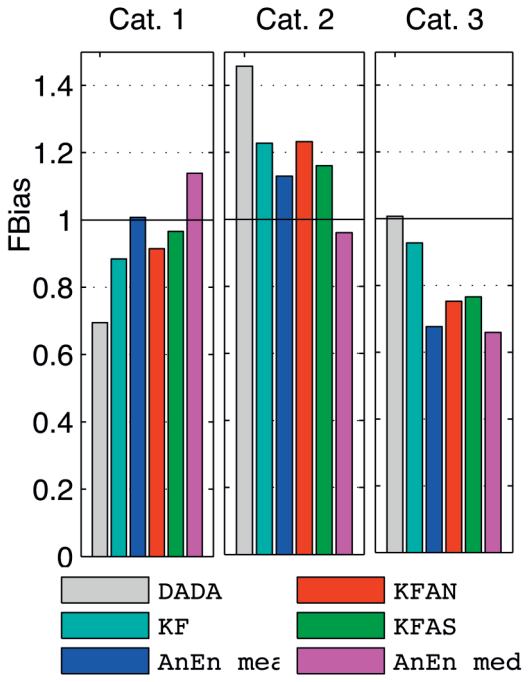

Figure 6. Frequency bias (Fbias) for 6 different deterministic forecasts averaged for 14 locations during year 2012. The Fbias is calculated for 3 different categories, so that Category 1 represents low (up to the 50th percentile), Category 2 moderate (between the 50th and the 90th percentile) and category 3 high wind speed (over the 90th percentile).

Slika 6. Učestalost pristranosti (Fbias) za 6 ispitanih deterministčkih prognoza usrednjenih za 14 lokacija tijekom 2012. Fbias je izračunat za 3 kategorije tako da kategorija 1 predstavlja slabi (do 50. percentila), kategorija 2 umjereni, a kategorija 3 jaki vjetra (iznad 90. percentila).

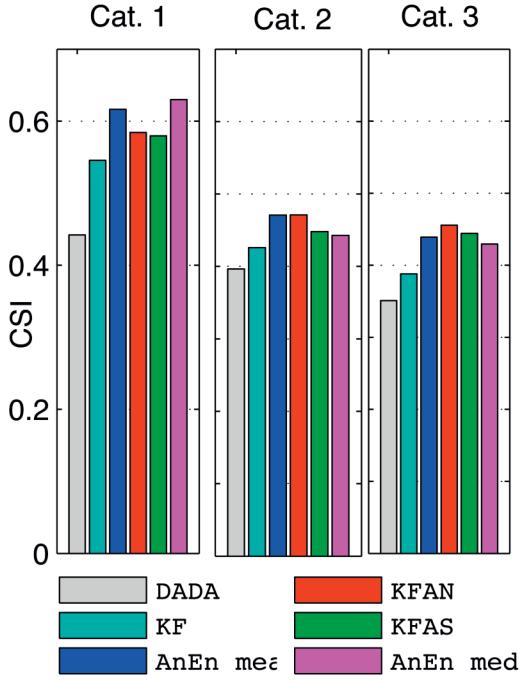

Figure 7. Critical success index (CSI) for 6 different deterministic forecasts averaged for 14 locations during year 2012. CSI is calculated for 3 different categories, so that Category 1 represents low (up to the 50th percentile), Category 2 moderate (between the 50th and the 90th percentile) and category 3 high wind speed (over the 90th percentile).

Slika 7. Kritični indeks uspješnosti (CSI) za 6 ispitanih deterministčkih prognoza usrednjenih za 14 lokacija tijekom 2012. Fbias je izračunat za 3 kategorije tako da kategorija 1 predstavlja slabi (do 50. percentila), kategorija 2 umjereni, a kategorija 3 jaki vjetra (iznad 90. percentila).

Table 1. Mean values of PCC for the DADA NWP model and the five different post-processing methods applied to it at 14 locations in Croatia during year 2012.

Tablica 1. Usrednjene vrijednosti PCC za DADA NWP model i 5 različitih post-procesnih metoda koje ovaj model koriste, za 14 lokacija u Hrvatskoj tijekom 2012.

	DADA	KF	AnEn mean	KFAN	KFAS	AnEn median
PCC	0.629	0.698	0.774	0.759	0.737	0.769

behavior as model data suggests that lack of radiation processes in this simplified model could only be partially compensated by any of these post-processing techniques. Nevertheless, there are some significant improvements due to post-processing, especially for AnEn mean and KFAN method. These two methods show similar errors, with KFAN having less bias, and AnEn mean having slightly better correlation.

Categorical verification showed similar results: AnEn mean has the highest PCC, and all of the analog based methods show higher values of PCC than KF and particularly than the DADA model (Table 1). Frequency bias shows that both the model and the post-processing methods predict medium wind speed too often (Figure 6). Low winds are underforecasted by the model, but post-processing improves that. On the other hand, model predicts higher wind speeds almost as often as they actually happen, while post-processing leads to underestimation of those cases. It can be concluded that post-processing reduces bias for common wind speeds and underestimates the frequency of rare ones. Again, the methods that include Kalman filtering provide the best results. Although less often predicted, strong wind speed category is predicted more accurately by post-processing methods, especially by the KFAN method. That results in a reduction of false alarms. All AnEn methods have similar CSI, better than KF and the DADA model (Figure 7). The different CSI values for different categories are likely a consequence of their sensitivity to climatology.

5.2. Probabilistic forecasting

Two probabilistic forecasts produced using the same dataset and training period are compared for the probabilistic prediction of wind speed exceeding 5 ms^{-1}. Statistical consistency measures (spread diagram and rank histogram) are only possible for the ensemble (AnEn) forecasts, so these results are not compared to LR.

The attributes diagram shows that wind speed exceeding 5 ms^{-1} occurs about 13 % of the cases, so it is a relatively rare event (Figure 8). The AnEn forecasting seems to be somewhat more reliable than LR, although both of them exhibit a good degree of reliability. For the smallest forecast probabilities (0-0.2 bin), both methods provide a slight overestimation of the observed relative frequency. Above that value, AnEn slightly overestimates, while LR underestimates the observed relative frequency. The forecast relative frequency for every bin reveals good tendency to predict extreme probabilities, with the majority occurring in the 0.0 - 0.2 probability range. The AnEn and LR forecasts are similarly sharp, except for AnEn (LR) predicts the highest (smallest) forecast probability bin a bit more often. Combined with good reliability, this is a trustworthy sharpness (i.e., does not result in overconfidence).

Since the attributes diagram is plotted considering all lead times, to better understand the performance over different lead times the BSS and its components are shown on Figure 9. Uncertainty term is the highest around noon, when wind speeds are the highest and climatological probability for wind speed to exceed 5 ms^{-1} is the highest (closer to 0.5). Both AnEn and LR forecasts show similar and very good ability to dicern subsample forecast periods with relative frequencies of the event that are different from each other. Resolution attribute is higher for AnEn forecasting than for LR forecasting for event and locations tested.

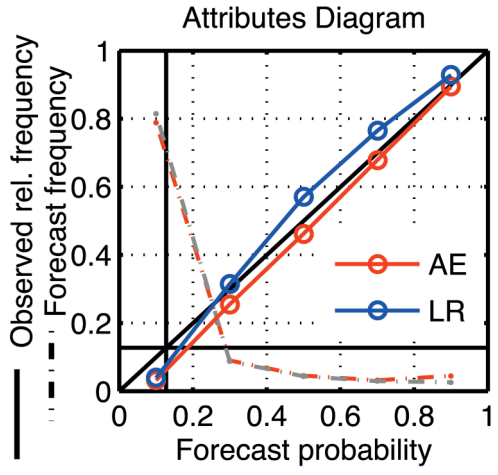

Figure 8. The attributes diagram of AnEn and LR probabilistic prediction of 10-m wind speed greater than 5 ms^{-1} at 14 locations during 2012.

Slika 8. Dijagram atributa za AnEn i LR probabilističku prognozu vjerojatnosti da brzina vjetra premaši 5 ms^{-1} tijekom 2012. na 14 postaja u Hrvatskoj.

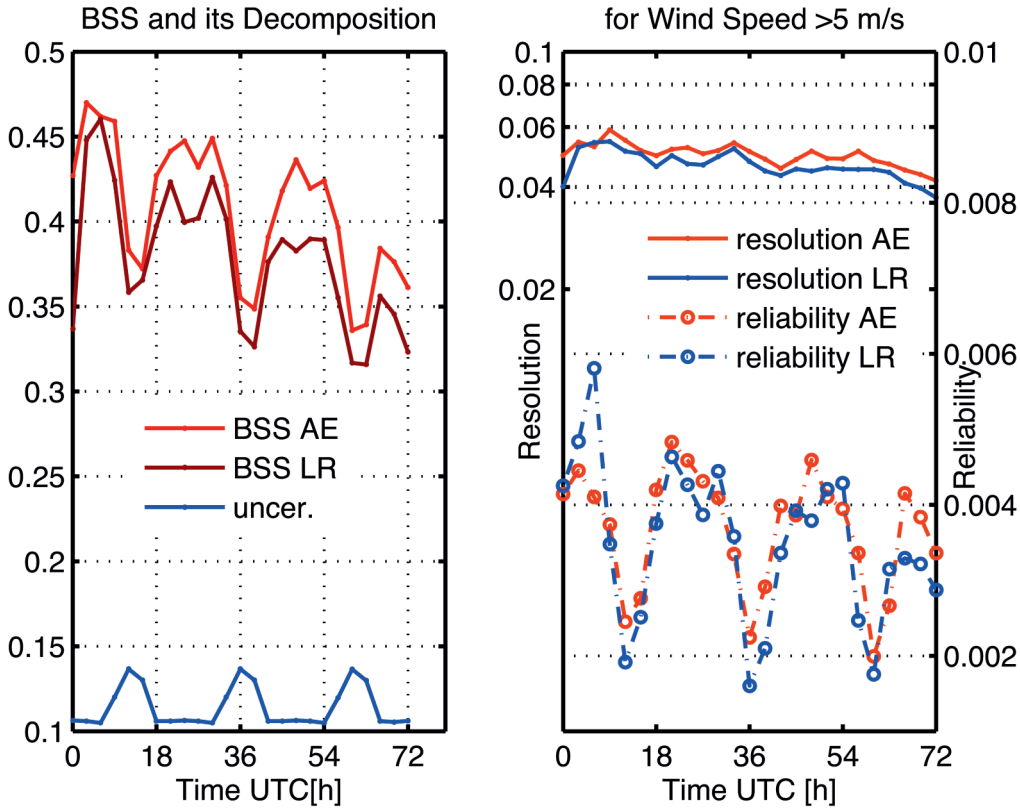

Figure 9. The Brier skill score and its decomposition to uncertainty, reliability and resolution, depending on lead time, for AnEn and LR probabilistic prediction of 10-m wind speed greater than 5 ms^{-1} at 14 locations during 2012..

Slika 9. Brierova mjera uspješnosti i njena dekompozicija na neizvjesnost, pouzdanost i rezoluciju u odnosu na nastupno vrijeme prognoze za AnEn i LR probabilističku prognozu vjerojatnosti da brzina vjetra premaši 5 ms^{-1} tijekom 2012 na 14 postaja u Hrvatskoj.

There is a small reduction of resolution for greater lead times, as expected. Reliability values show minimums around noon (highest wind speeds and uncertainty). It increases (while wind speeds and uncertainty decrease) for larger lead times, especially for LR forecasting. The latter result, together with the attributes diagram indicates that AnEn is a more reliable forecast when predicting high probabilities, while LR reliability is more consistent with lead time, mostly due to the 0.2-0.4 bin. Finally, the BSS results show better values during nights than around noon and that the AnEn forecasting system has a better relative skill than LR. These results might not seem intuitive at first glance, but are consistent with the previous analysis. Even though LR is more reliable for greater lead times, that difference in reliability term between AnEn and LR forecasting is around an order of magnitude smaller than the difference in resolution term. Also, the BSS minimums around noon are consequence of increased uncertainty, so they are mostly affected by the climatological distribution of the observations and not the forecasts. Without normalization with uncertainty (that would result in the Brier score), score minimums would be more consistent with reliability maximums and resolution minimums.

Figure 10 consists of ROC skill score results for AnEn and LR forecasting depending on lead time, dispersion diagram and rank histogram for AnEn forecasting. Results regarding ROC skill score show that the AnEn performs better than LR regarding forecast ability to discriminate between two possible outcomes (when wind speed did or did not exceed 5 ms^{-1}). Since ROCSS depends on resolution (and not on reliability), this is an expected result. Better result for AnEn forecasting system

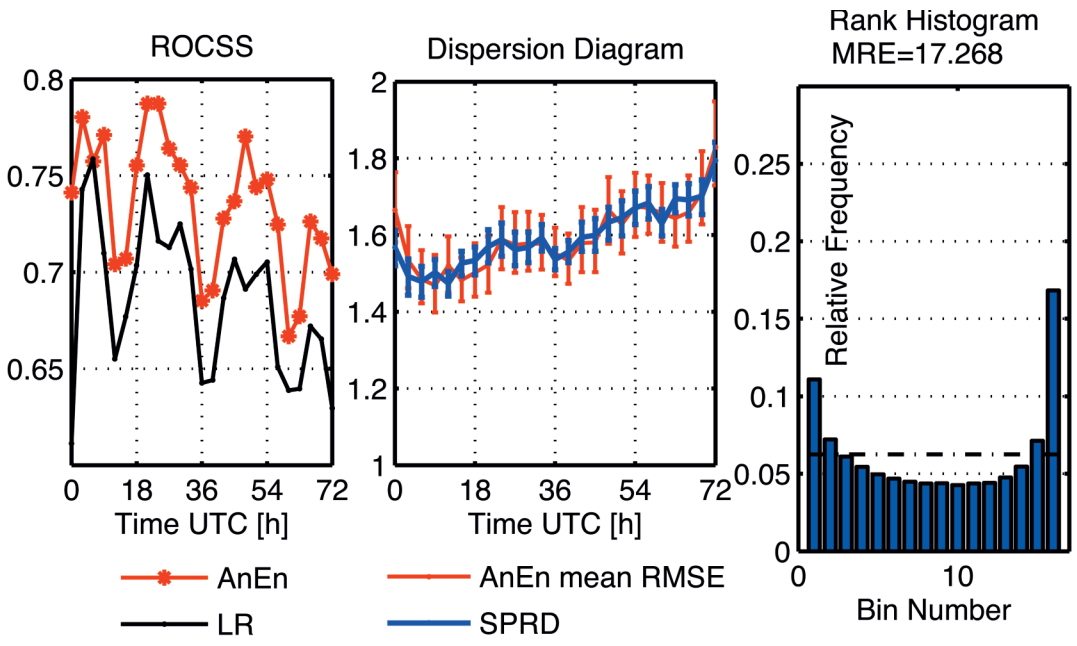

Figure 10. Left: ROC skill score result for AnEn and LR forecasting depending on lead time. Middle: Dispersion diagram for AnEn forecasting. Right: Rank histogram for AnEn forecasting. All results refer to 14 locations in Croatia during year 2012.

Slika 10. Lijevo je prikazana ROC mjera uspješnosti u odnosu na nastupno vrijeme prognoze za AnEn i LR probabilističku prognozu tijekom 2012. na 14 postaja u Hrvatskoj. U sredini i desno za iste podatke prikazan je dijagram raspršenja i histogram ranga.

is probably due to LR's difficulty to fit a dependable regression line in case of the event is rarely observed within the training dataset, as it is the case here.

Statistical consistency is examined via dispersion diagram and rank histogram, only for AnEn forecasting. The LR forecasting is not included because in its standard formulation (the one adopted in this study) it does not provide an actual ensemble, just a probability for a given event threshold. The rank histogram of AnEn is U-shaped. That result indicates either slight under-spread condition that could not be corrected with increasing the number of ensemble members, or the ensemble is sampling a population with some combination of conditional biases (as noted in Hamill 2001). These results are consistent with the results presented in Delle Monache et al. (2013) and the time-lagged ensemble forecasting in Lu et al. (2007). The MRE value shows that it is more probable for the observed value to be higher than any of the ensemble members, than to be lower than any AnEn member. This result includes all lead times, but it is expected that the histogram flattens with increasing lead time (Candile and Talagrand 2005). To see how statistical consistency varies with lead time, a dispersion diagram can be plotted. Dispersion diagram shows that the mean square error of the ensemble mean matches the average ensemble variance, suggesting that AnEn is properly dispersive.

6. CONCLUSIONS

This study compares different post-processing methods based on a historical data set of deterministic 2-km dynamical adaptation model (DADA) 00 UTC runs and verifying observations of 10-m wind speed. For deterministic post-processing of 10-m wind speed all of the methods tested reduce forecast error and bias compared to the DADA model, while at the same time they are improving correlation. The largest bias reduction is obtained with methods that use Kalman filtering, while analog-based methods show the highest correlation and the smallest root-mean-square-error. Analog-based post processing methods show improvement in

forecasting high wind speed also, even though they are forecasted too rarely. Overall, the best results are for the analog ensemble (AnEn) mean and KF forecasting applied to the mean of AnEn (KFAN).

Regarding probabilistic post-processing of wind speed forecasts, probability for wind speed to exceed the analyzed threshold (5 m/s) is better predicted with the analog-based method than with logistic regression (LR) post-processing method. Both forecast systems were similarly sharp, but AnEn has better resolution and discrimination attributes than LR. It is also more reliable when predicting high probability of event occurrence. The LR forecasting is somewhat more reliable for greater lead times than AnEn , but that is likely due to the low probability class that is close to climatological forecasting. Statistical consistency was also tested. The rank histogram for AnEn shows a mild U-shape that suggests slight underspread or conditional bias, while dispersion diagram shows satisfactory dispersion.

Based on these results, analog ensemble method shows to be useful for post-processing of ALADIN numerical weather forecasts. For deterministic post-processing the best results are for AnEn mean and KFAN forecasting. Furthermore, AnEn method applies not only to the improvement of accuracy of deterministic forecast, but even more to providing reliable forecast uncertainty information for locations where measurements do exist. Therefore, AnEn method is being additionally tested in quasi-operational setting to assess its potential as a component of the numerical wind prediction system.

ACKNOWLEDGEMENTS

This study was supported by the grant IPA2007/HR/16IPO/001-040507 through the WILL4WIND project (www.will4wind.hr).

REFERENCES

ALADIN International Team, 1997: The ALADIN project: Mesoscale modelling seen as a basic tool for weather forecasting and atmospheric research. WMO Bull., 46, 317-324.

Alessandrini, S., L. Delle Monache, S. Sperati, G. Cervone, 2015a: An analog ensemble for short-term probabilistic solar power forecast. *Applied Energy*, 157, 95-110.

Alessandrini, S., L. Delle Monache, S. Sperati, J. Nissen, 2015b: A novel application of an analog ensamble for short-term wind power forecast. *Renewable Energy*, 76, 768-781.

Bajić, A., 2003: Očekivani režim strujanja vjetra na autocesti Sv. Rok (jug) - Maslenica. *Građevinar*, 55, 149-158.

Bajić, A.,S. Ivatek-Šahdan, K. Horvath, 2007: Spatial distribution of wind speed in Croatia obtained using the ALADIN model. *Cro. Met. J.* 42, 67-77.

Bajić, A.,S. Ivatek-Šahdan, Z. Žibrat, 2008: ANEMO-ALARM iskustva operativne primjene prognoze smjera i brzine vjetra. *GIU Hrvatski cestar.* 109-114.

Candile, G., O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. *Q.J.R. Meteor. Soc.*, 131, 2131-2150.

Delle Monache, L., T. Nipen, X. Deng, Y. Zhou, R. B. Stull, 2006: Ozone ensemble forecasts: 2. A Kalman filter predictor bias-correction. *Journal of Geophysical Research* 111, D05308.

Delle Monache, L., J. Wilczak, S. McKeen, G. Grell, M. Pagowski, S. Peckham, R. Stull, J. McHenry, J. McQueen, 2008: A Kalman-filter bias correction of ozone deterministic, ensemble-averaged, and probabilistic forecasts. *Tellus B*, 60, 238-249.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, R. Stull, 2011: Kalman filter and analog schemes to post-process numerical weather predictions. *Mon. Wea. Rev.* 139, 3554-3570.

Delle Monache, L., T. Eckel, D. Rife, B. Nagarajan, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.* 141, 3498-3516.

Djalalova, I., L. Delle Monache, J. Wilczak, 2015: PM2.5 analog forecast and Kalman filter post-processing for the Community Multiscale Air Quality (CMAQ) model. *Atmospheric Environment* 108, 76-87.

Drosdowsky, W., 1994: Analog (nonlinear) forecasts of the Southern Oscillation index time series. *Weather and Forecasting*, 9, 78-84.

Eckel, F. A., C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting*, 20, 328-350.

Esterle, G. R., 1992: Adaptive, self-learning statistical interpretation system for the central Asian region. *Ann. Geophys.*, 10, 924-929.

Gao, L., H. Ren, J. Li, J. Chou, 2006: Analogue correction method of errors and its application to numerical weather prediction. *Chin. Phys.*, 15, 882-889.

Geleyn J.-F., 1988: Interpolation of wind, temperature and humidity values from model levels to the height of measurement. *Tellus*, 40A, 347-351.

Hamill, T.M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, 129, 550-560.

Hamill, T. M., J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, 134, 3209-3229.

Hopson, T. M., P. J. Webster, 2010: A 1Ð10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003-07. *J. Hydrol.*, 11, 618-641.

Horvath, K., S. Ivatek-Šahdan, B. Ivančan-Picek, V. Grubišić, 2009: Evolution and structure of two severe cyclonic bora events: contrast between the northern and southern Adriatic. *Weather and forecasting* 24, 946-964.

Horvath, K., A. Bajić, S. Ivatek-Šahdan, 2011: Dynamical Downscaling of Wind Speed in Complex Terrain Prone to Bora-Type Flows. *Journal of Applied Meteorology and Climatology*, 1676-1691.

Ivatek-Šahdan, S., M. Tudor, 2004: Use of high-resolution dynamical adaptation in operational suite and research impact studies. *Meteorol Zeitschrift* 13 (2), 99-108.

Ivatek-Šahdan, S., B. Ivančan-Picek, 2006: Effects of different initial and boundary conditions in ALADIN/HR simulations during MAP IOPs. *Meteorol Zeitschrift* 15, 187-197.

Jolliffe I. T., D.B. Stephenson, 2011: Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley, 292 pp.

Junk, C., L. Delle Monache, S. Alessandrini, L. von Bremen, G. Cervone, 2015: Predictor-weighting strategies for probabilistic wind power forecasting with an analog ensemble. *Meteorol Zeitschrift* (accepted).

Juras, J., Z. Pasarić, 2006: Application of tetrachoric and polychoric correlation coefficients to forecast verification. *Geofizika*, 23, 59-81.

Kalman, R. E., 1960: A new approach to linear filtering and prediction problems. *J. Basic Eng.*, 82, 35-45.

Klausner, Z., H. Kaplan, E. Fattal, 2009: The similar days method for predicting near surface wind vectors. *Meteor. Appl.*, 16, 569-579.

Lu, C., H. Yuan, B. E. Schwartz, S. G. Benjamin, 2007: Short-Range Numerical Weather Prediction Using Time-Lagged Ensembles. *Weather and Forecasting*, 22, 580-595.

Murphy, A. H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8, 281-293.

Panziera, L., U. Germann, M. Gabella, P. V. Mandapaka, 2011: NORA—Nowcasting of orographic rainfall by means of analogues. *Quart. J. Roy. Meteor. Soc.*, 137, 2106-2123.

Ren, H., J. Chou, 2006: Analogue correction method of errors by combining statistical and dynamical methods. *Acta Meteor. Sin.*, 20, 367-373.

Ren, H., J. Chou, 2007: Strategy and methodology of dynamical analogue prediction. *Sci. China Ser. D: Earth Sci.*, 50, 1589-1599.

Stanešić, A., 2011: Assimilation system at DHMZ: development and first verification results, *Cro. Meteorol. J.*, 44/45, 3-17.

Talagrand, O., R. Vautard, B. Strauss, 1997: Evaluation of probabilistic prediction systems. Proc. Workshop on Predictability, Reading, United Kingdom, ECMWF, 1-25.

Tudor, M., S. Ivatek-Šahdan, 2002: The MAP-IOP 15 case study. *Cro. Met. J.* 37, 1-14.

Tudor, M., S. Ivatek-Šahdan, A. Stanešić, K. Horvath, A. Bajić, 2013: Forecasting weather in Croatia using ALADIN numerical weather prediction model. *Climate Change and Regional/Local Responses*, InTech, 247 pp., 59-88.

Van den Dool, H. M., 1989: A new look at weather forecast through analogs. *Mon. Wea. Rev.*, 117, 2230-2247.

Vanvyve, E., L. Delle Monache, D. Rife, A. Monaghan, J. Pinto, 2015. Wind resource estimates with an analog ensemble approach. *Renewable Energy*, 74, 761-773.

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. 3rd ed., Academic Press, 676 pp.

Wu, W., and Coauthors, 2012: Statistical downscaling of climate forecast system seasonal predictions for the southeastern Mediterranean. *Atmos. Res.*, 118, 346-356.

Xavier, P. K., B. N. Goswami, 2007: An analog method for real-time forecasting of summer monsoon subseasonal variability. *Mon. Wea. Rev.*, 135, 4149-4160.

Žagar, M., J. Rakovec, 1999: Small-scale surface wind prediciton using dynamical adaptation. *Tellus*, 51A, 489-504.