Verification of Gridded Wind Forecasts in Complex Alpine Terrain: A New Wind Verification Methodology Based on the Neighborhood Approach

GREGOR SKOK

Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

VERONIKA HLADNIK

Slovenian Environment Agency, Ljubljana, Slovenia

(Manuscript received 13 December 2016, in final form 17 October 2017)

ABSTRACT

A novel wind verification methodology is presented and analyzed for six surface wind cases in the greater Alpine region as well as an idealized setup. The methodology is based on the idea of the fractions skill score, a neighborhood-based spatial verification metric frequently used for verifying precipitation. The new score avoids the problems of traditional nonspatial verification metrics (the “double penalty” problem and the failure to distinguish between a “near miss” and much poorer forecasts) and can distinguish forecasts even when the spatial displacement of wind patterns is large. Moreover, the time-averaged score value in combination with a statistical significance test enables different wind forecasts to be ranked by their performance.

1. Introduction

In recent years, new verification methods have been developed for verifying spatial forecasts of precipitation (e.g., the ICP project at www.ral.ucar.edu/projects/icp; Brown et al. 2012; Ebert 2008; Rossa et al. 2008; Gilleland et al. 2009, 2010; Gilleland 2013a). These spatial verification metrics try to address the shortcomings of the traditionally used nonspatial precipitation verification metrics that suffer from several problems.

A common problem encountered in many forecasts is an offset in the predicted position of a weather event relative to where it actually occurred. When using a nonspatial verification metric, such a forecast incurs a “double penalty” since it is penalized for predicting an event where it did not occur, and again for failing to predict the event where it did actually occur. Another problem is that nonspatial scores do not distinguish between a “near miss” and much poorer forecasts. These problems tend to provide results that are inconsistent with a subjective evaluation of

the forecast made by a human analyst. The need to reconcile subjective evaluations of forecast quality with objective measures of forecast accuracy has been the main motivation in researching and developing diagnostic methods for spatial verification that more adequately reflect their worth (Brown et al. 2012).

Although efforts to develop spatial verification metrics for precipitation have been under way for some time, very little improvement has been made regarding wind verification methods. This shortcoming was recognized and addressed by the ongoing Mesoscale spatial forecast Verification Intercomparison over Complex Terrain (MesoVICT) project (Dorninger et al. 2013) that, besides focusing on precipitation verification in complex terrain and ensemble forecasts, lists progress in wind verification as one of its primary goals.

Horizontal wind verification poses a specific problem since wind and precipitation fields are fundamentally different: the former is a vector field while the latter is a scalar field. The problem is usually avoided by converting the wind field into a scalar field by separately verifying the wind speed and wind direction or separately verifying the two wind components, making it more difficult to interpret the results.

At higher altitudes, horizontal wind is usually a relatively smooth variable (with the exception of wind near

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/MWR-D-16-0471.s1>.

Corresponding author: Gregor Skok, gregor.skok@fmf.uni-lj.si

areas of strong convection or strong wind shear), which changes only gradually while, near the surface, the wind speed and direction might alter dramatically at smaller scales in response to changes in topography. The topography in a numerical model is usually smoothed because of the limitations of the model's resolution. This smoothed topography can result in different wind patterns in which the winds might have a different speed or direction or be displaced due to the incorrect position of the topographic slopes. The spatial displacement of wind patterns also commonly happens at all altitudes in global models at synoptic scales with large forecast times (e.g., 9 days) since the synoptic-scale features, such as cyclones and frontal systems, might be significantly displaced. Thus, the spatial displacement error of wind is a common occurrence, and a double-penalty problem will arise, much like in the case of precipitation.

The fractions skill score (FSS; Roberts and Lean 2008; Roberts 2008) is a popular spatial verification metric used for verifying precipitation. It uses the concept of a neighborhood to alleviate the requirement that the event should be forecasted at exactly the correct position. The FSS can provide a truthful assessment of displacement errors and forecast skill and also enables forecasts of different resolutions to be compared against a common spatial truth (e.g., radar rainfall analyses) in such a way that high-resolution forecasts are not penalized for representativeness errors that arise from the double-penalty problem (Mittermaier and Roberts 2010; Mittermaier et al. 2013). The score can also be calculated very efficiently at low computational cost (Faggian et al. 2015). On the other hand, the score value can be influenced by the bias in the forecast, the orientation of the displacement, and the existence of a nearby domain border (Mittermaier and Roberts 2010; Skok 2015, 2016; Skok and Roberts 2016). The score, as originally defined, can analyze only scalar variables. Here an attempt is made to extend the score to analyze wind vector fields in a meaningful way. The new score is calculated and analyzed for cases in the MesoVICT database as well as an idealized setup.

Section 2 provides a definition of the new score and an evaluation of the score for a simple idealized case. Section 3 focuses on an analysis of the MesoVICT near-surface wind over the greater Alpine region, while section 4 provides the conclusions.

2. The new score

a. Definition of the FSS_{wind}

Since the score is based on the idea of the FSS, the new score is named the wind fractions skill score, and

denoted as FSS_{wind} . The score is calculated using the following steps. First, a number of wind classes have to be defined using some predefined rules. The total number of classes is denoted as P . A single class can be present at each location. It is important that the class definitions cover the whole phase space of possible wind values (i.e., that any wind vector can be assigned to a wind class). Assuming no data are missing, a wind class will be defined at every grid point inside the domain.

Next, the field is decomposed into a number of binary fields according to classes. The binary fields can only have values of 0 or 1. The value of 0 represents locations where a certain wind class is not present, while the value of 1 represents locations where a certain wind class is present. For each input wind vector field, there are P binary fields. Once the binary fields are constructed, the fractions are calculated in the same way as for the original FSS (see Roberts and Lean 2008), but for every wind class separately. Fractions can have values between 0 and 1 and represent a portion of the nonzero area inside a certain neighborhood. The neighborhood is assumed to be a square of size n points, thus consisting of n^2 grid points (e.g., a neighborhood of size $n = 5$ consists of $5 \times 5 = 25$ grid points). The n is assumed to be an odd integer, and the area outside the domain is assumed to contain no wind class. Calculating the fractions is the computationally most intensive task in the calculation of the FSS_{wind} value (especially for large neighborhoods). However, since the fractions are calculated in the same way as for the original FSS, and assuming the domain is rectangular and a square neighborhood is used, the calculation can be done computationally very efficiently using the idea of summed fields presented by Faggian et al. (2015). An alternative option is to use the fast Fourier transform (FFT) along with the convolution theorem to calculate the fractions (e.g., by using the “Smoothie” R software package; Gilleland 2013b). In this case, other neighborhood shapes could be used (e.g., circular or Gaussian). However, the computational cost of the Faggian approach is proportional to M (with M being the total number of grid points in the domain); while the computational cost of the FFT approach is proportional to $M \times \log M$, making the Faggian approach significantly faster than the FFT.¹

Once the fraction values are calculated, the FSS_{wind} can be calculated as

¹ The R software source code for calculating the FSS_{wind} value, which already includes all the computational tricks for fast calculation, is available upon request from the corresponding author.

$$\text{FSS}_{\text{wind}} = 1 - \frac{\sum_k \sum_{i,j} [O_k(i,j) - M_k(i,j)]^2}{\sum_k \sum_{i,j} O_k(i,j)^2 + \sum_k \sum_{i,j} M_k(i,j)^2}, \quad (1)$$

where $O_k(i, j)$ represents the fraction value for observations for wind class k at location i, j while $M_k(i, j)$ represents the same thing for the forecast. In comparison with the original FSS, there is an additional sum over all the wind classes ($k = 1, \dots, P$). The range of the FSS_{wind} is between 0 and 1, with 1 indicating a perfect forecast and 0 indicating the worst possible forecast. A FSS_{wind} value of 1 occurs when the wind class fractions for the two fields are identical at all grid points in the domain. For example, if a FSS_{wind} value of 1 occurs at the smallest neighborhood size ($n = 1$) then the position and extent of the wind classes in the forecast perfectly matches with the position and extent of the wind classes in the observations. On the other hand, if an FSS_{wind} value of 0 occurs at $n = 1$ then all grid points in the domain would have different wind classes in the forecast and observations.

The FSS_{wind} value depends on the neighborhood size. Similarly to the original FSS, the FSS_{wind} value usually increases with the neighborhood size. An increase in the FSS_{wind} value happens since, when the neighborhood size is increased, the regions of the same wind class are allowed to be ever more spatially displaced in the forecasts. Thus, the requirement that the correct wind class in the forecast be located exactly in the correct location is relaxed (thus avoiding the problems of non-spatial verification metrics). This property can be considered one of the most advantageous properties of the new score.

Once the neighborhood is large enough, the asymptotic FSS_{wind} value is reached. This value is denoted as $\text{FSS}_{\text{wind}}^{\text{asy}}$ and will always be reached (assuming a square neighborhood and rectangular domain) when $n \geq 2N_x + 1$, where N_x denotes the length of the largest domain side. For such large neighborhoods, the fractions inside the domain are guaranteed to be the same at all locations within the domain and the asymptotic value can be derived from Eq. (1) as

$$\text{FSS}_{\text{wind}}^{\text{asy}} = 1 - \frac{\sum_k (f_k^O - f_k^M)^2}{\sum_k (f_k^O)^2 + \sum_k (f_k^M)^2}, \quad (2)$$

where f_k^O denotes the frequency of wind class k in the observations and f_k^M is the frequency of the same wind class in the forecast. The frequency is defined as the number of grid points in a certain wind class divided by the number of all grid points in the domain. Thus, it

follows that $\text{FSS}_{\text{wind}}^{\text{asy}}$ will be equal to 1 only if the frequency of the corresponding wind classes in the observations and the forecast are the same ($f_k^O = f_k^M$, for $k = 1, \dots, P$). This property shows that the FSS_{wind} value will depend not only on the degree of spatial matching, but also on the bias in the two wind fields (represented by the difference in frequencies of corresponding wind classes).

It is important to highlight that, contrary to the original FSS, the FSS_{wind} value will always be defined. Namely, the FSS value cannot be calculated for cases when there are no rainy grid points in the domain (since a division with zero occurs), which can frequently occur in the real world if a smaller domain is used. On the other hand, the FSS_{wind} value can always be calculated since it is guaranteed that a wind class is defined at every location inside the domain (the only exception being if the whole domain contains only missing data). This makes the FSS_{wind} more stable than the original FSS.

It is worth highlighting that the FSS_{wind} provides a single value as the output. This property is important since a single output value enables different forecasts to be ranked by their performance. We believe the ability to rank forecasts by performance is an important property of any verification score. Contrary to the FSS_{wind} , the binary fields of each wind class could also be evaluated separately by using the original FSS, thus providing information about each wind class separately. However, if one of the wind classes happens to be totally missing from the fields, the FSS value cannot be calculated for that class. Moreover, even if the FSS value for each class could always be calculated (i.e., no division by zero would occur), the information about different classes would have to be combined together into a single metric for the ranking to become possible. It is unclear how this can be achieved in the case of the original FSS, at least not in a more elegant and more compact way than how it is already done in Eq. (1).

b. A demonstration of the new score's spatial aspect

Traditional nonspatial grid-scale metrics only compare values at the same locations. This presents a problem since a common problem in many forecasts is an offset in the predicted position of a weather event relative to where it actually occurred. As previously mentioned, nonspatial scores do not distinguish between a near miss and much poorer forecasts. However, since the FSS_{wind} is a spatial verification metric it thus avoids this problem. To demonstrate the spatial aspect of the new score, we use a simple idealized case and analyze the forecast using the FSS_{wind} and a simple nonspatial metric.

FIG. 1. A demonstration of the spatial aspect of the new score using an idealized experiment. (a),(b) The layouts of the observed and forecasted wind fields. (c) The dependence of the score value on the displacement of the veered wind classes in the forecasted field. The FSS_{wind} value is calculated for a neighborhood size of 201 grid points.

For the nonspatial metric, we devise a simple root-mean-square error metric that is adapted for use with wind and denoted as $RMSE_{\text{wind}}$. The definition is

$$RMSE_{\text{wind}} = \sqrt{\frac{1}{M} \sum_{i,j} |\mathbf{v}_{\text{obs}}(i,j) - \mathbf{v}_{\text{fcs}}(i,j)|^2}, \quad (3)$$

where M is the total number of grid points in the domain, while $\mathbf{v}_{\text{obs}}(i,j)$ and $\mathbf{v}_{\text{fcs}}(i,j)$ are the observed and forecasted wind vectors at location (i,j) . The sum over i and j is over all the grid points in the domain. The absolute operator represents the Euclidian vector norm. The $RMSE_{\text{wind}}$ is a nonspatial metric since it only compares the observed and forecasted wind vectors at collocated grid points.

The values of the two scores are calculated for an idealized case shown in Fig. 1. The idealized setup comprises a domain consisting of 500×500 grid points, where the prevalent wind of 1 m s^{-1} comes from the right side but veers toward the top and bottom of the domain while maintaining speed. The veering happens in a smaller elliptically shaped region (major and minor ellipse axes are 360 and 100 grid points). This kind of veering can be the result of the prevailing flow impinging upon a mountain slope, which obstructs the incoming flow. In the forecast, the area where the wind veers is spatially displaced along the direction of the prevailing flow. The displacement can happen because of the unrealistic representation of orography in the model, which is usually too smooth.

Figure 1c shows how the values of $RMSE_{\text{wind}}$ and FSS_{wind} depend on the size of the displacement. For this particular case, three classes are defined and used for the calculation of FSS_{wind} : the prevailing wind class with a flow toward the right and the two classes for the veered

flows. At zero displacement, both scores provide the perfect value: zero for $RMSE_{\text{wind}}$ and one for FSS_{wind} . Once the displacement increases, the value of both scores becomes worse (larger for $RMSE_{\text{wind}}$ and smaller for FSS_{wind}). Yet, once the displacement reaches 100 grid points (when the two elliptic regions stop overlapping), the increase of $RMSE_{\text{wind}}$ stops. This illustrates the problem of nonspatial scores that cannot distinguish between a near miss and much poorer forecasts. On the other hand, the FSS_{wind} value continues to decrease well beyond the displacement of the 100 grid points. This highlights the new score's ability to distinguish the forecasts even when the spatial displacement is large.

At the same time, it is important to note that, while the FSS_{wind} is indeed sensitive to large spatial displacements, the FSS_{wind} values in the idealized case only span the range between 1 and approximately 0.9. The existence of a relatively high minimum FSS_{wind} value of 0.9 might be surprising, but eventually highlights another property of the score, namely, that the FSS_{wind} value in the idealized case is mainly determined by the prevailing wind class. Since the prevailing wind class covers the majority of the domain in both fields (about 89% of the domain area), the score value will always tend to be high because of the good spatial match of this wind class. Thus, the effect of the spatial displacement error of the other wind class on the score value is limited. The influence of a certain class on the overall score value depends mainly on the class frequency, meaning the score value will be under the greatest influence of the most frequent class in the fields. This also means that if some wind class is rare (e.g., occurs only occasionally or covers only a very small portion of the domain), it will not significantly influence the score value. It is important to keep this property in mind when interpreting the results.

3. Analysis of the MesoVICT surface wind

a. Description of the data

The behavior of the new score is tested on the MesoVICT dataset. MesoVICT (Dorninger et al. 2013) is a community project aimed at the development and analysis of verification methods. One of its primary aims is the development of verification methods for wind forecasts. The project also provides a community test bed where common datasets are available.

The MesoVICT test bed dataset consists of several cases. The dataset includes gridded data for surface wind obtained via the Vienna Enhanced Resolution Analysis (VERA) as well as forecasts provided by different models. The data have hourly resolution and are available on an 8-km regular grid in the Alps and surrounding area (Fig. 2). The analysis is performed for six MesoVICT cases (Table 1) on a rectangular domain, consisting of 88×133 grid points.

The VERA surface wind was obtained with use of the VERA scheme (Steinacker et al. 2000) that performs an interpolation of sparsely and irregularly distributed observations to a regular grid in mountainous terrain. The algorithm is based on a thin-plate spline approach and the analysis is independent of any model data. The resolution of the VERA analysis surface wind data is 8 km. Data from two models are usually available: forecasts from the 2.2-km Swiss model COSMO2 (denoted as CO2_00 and CO2_06 for setups initialized at 0000 and 0600 UTC) provided by MeteoSwiss, and forecasts from the 2.5-km Canadian high-resolution model Global Environmental Multiscale Limited-Area Model (GEMLAM) (denoted as CMH_06) provided by Environment Canada. The wind fields of the selected models and the VERA analysis are provided as ASCII files, with the model fields interpolated onto the VERA analysis grid. More information about the cases, models, and VERA analysis is available in Dorninger et al. (2013).

b. Definition of wind classes and calculation of the score value

To calculate the FSS_{wind} , wind classes have to be defined. It is important to note that the definition of classes is partly subjective. As stated before, one of the main aims of the spatial verification methods is to bring the verification more in line with a subjective analysis made by a human analyst. The choice of classes will define how the score behaves and influence the results. The definition of classes should reflect what a user wants to verify. Since the near-surface wind in mountainous regions is heavily influenced by the orography, which can cause changes in direction and speed, it is reasonable to differentiate the wind classes

FIG. 2. The MesoVICT domain used for the FSS_{wind} analysis (black rectangle). The gray shading shows the topographic elevation.

according to the wind direction and speed. Figure 3a shows the wind rose averaged over all the MesoVICT cases for the VERA analysis over the whole domain. Overall, the wind coming from the southerly quadrant is most frequent with about 23% of wind being less than 1 m s^{-1} .

The simplest natural choice for differentiation according to wind direction is the four cardinal winds (American Meteorological Society 2017). It may also make sense to define a special wind class to represent calm/low wind, irrespective of the wind direction, since some types of widely used anemometers have problems estimating the wind speed or direction at low wind speeds. For example, medium-size cup anemometers require a minimum wind speed of around 1 m s^{-1} to work properly (Harrison 2014), which can be used as a threshold for the wind class.

We thus define five basic wind classes, as shown in Fig. 3b. Any wind with a speed below 1 m s^{-1} is defined as the calm/low wind class (regardless of the wind's direction). If the wind speed exceeds the 1 m s^{-1} threshold, the wind is classified into one of the four additional direction-dependent classes (northerly, southerly, easterly, and westerly). Each direction-dependent class spans an azimuthal range of 90° . Figure 3c shows the frequency of wind classes averaged over all the MesoVICT cases. Although some differences do exist, the frequencies are quite evenly spread between different classes (e.g., there is no single class that would totally dominate other classes or some class that would be almost totally absent).

The use of classes presented in Fig. 3b will result in a score evaluating the spatial matching of the areas of the

TABLE 1. Description of the MesoVICT cases used for the FSS_{wind} analysis.

	Period	Duration	Available models
Case 1 (core case)	20–22 Jun 2007	56 h	CO2_00, CMH_06
Strong convective developments north of the Alps followed by a cold front the next day. The cold air mass could not spill over the Alps.			
Case 2	18–21 Jul 2007	75 h	CO2_00, CMH_06, CO2_06
Stationary airmass boundary slightly north of the Alps, strong convective events along this airmass boundary. Winds are generally weak except in the surrounding area of the convective cells.			
Case 3	25–28 Sep 2007	75 h	CO2_00, CMH_06, CO2_06
A fast-moving cold front impinges the Alps from NW, cyclone development in the Gulf of Genoa causing severe convection south of the Alps (Venice–Mestre flood).			
Case 4	6–8 Aug 2007	56 h	CO2_00, CMH_06, CO2_06
A summery convective situation in the Alps par excellence. Ahead of a cold front, a squall line moves slowly over the Alps, leading to widespread convective events in large parts of the area. Strong and gusty winds are observed in the vicinity of the convective cells.			
Case 5	18 Sep 2007	18 h	CO2_00, CMH_06, CO2_06
Two cold fronts are passing the area north of the Alps. A specific detail of this case is that the first cold front makes it over the eastern Alps and initiates strong thunderstorms in the area of Slovenia. These thunderstorms are very stationary and can cause local flooding.			
Case 6	8–10 Jul 2007	56 h	CO2_00, CMH_06
In a moist and unstable subtropical air mass, convective events are frequently initiated and move slowly from west to east over the whole Alpine area.			

five basic wind classes. While the use of the five basic classes is reasonable in terms of surface wind in mountainous terrain, the classes might be defined differently for other cases. For example, when verifying upper-level jet streams, the direction of the wind might not be considered important and the focus could be solely on the magnitude of the wind speed. In this case, the classes could be defined only according to wind speed and the score would evaluate the spatial matching of the areas of different wind speeds only.

We advise the user to carefully consider how to define the wind classes. As stated before, the class definitions should cover the whole phase space of possible wind values (i.e., to make sure every wind vector can be assigned a wind class)—this will ensure the score value can always be calculated (i.e., no division with zero can occur). Moreover, it makes little sense choosing a definition with an overly large number of classes since

this does not reflect how a subjective analysis by a human analyst would be done. A human analyst would not subjectively decompose a wind field into 100 classes before visually estimating a spatial match of the corresponding class regions. It also does not make sense to define classes in such a way that a single class would totally dominate over all the other classes in the dataset (that a single class would cover almost the whole domain in both fields). If this were the case, the score value would always tend to be near-perfect since the overlap of the dominating class would be nearly perfect over the whole domain. When deciding on how to define the wind classes, we advise the user to always analyze the wind rose corresponding to the dataset. Once the class definition is chosen, the frequency of all classes should be checked in order to avoid a single class being overly dominant. To avoid discrimination of certain forecasts, all forecasts should always be verified over the same

FIG. 3. (a) Wind rose for VERA analysis wind averaged over all cases and the whole domain. (b) The definition of the five basic wind classes. (c) The class frequency according to the basic class definition averaged over all the MesoVICT cases.

FIG. 4. An example of conversion of the wind fields into wind class index fields. Shown is the MesoVICT case 3 at 0600 UTC 25 Sep 2007 using the CO2_00 model forecast: (a),(b) the two input near-surface wind vector fields and (c),(d) the resulting wind class index fields according to the five basic wind classes from Fig. 3b.

domain using an identical grid and the same class definitions.

Figures 4–6 demonstrate the process of calculating the score value. Figure 4 shows the original surface wind vector fields for a selected example and the resulting wind class index fields. The index fields are obtained using the five basic classes from Fig. 3b. Figure 5 shows binary wind class fields for the same example. The binary fields are used to calculate the fractions needed by Eq. (1). These fractions can have values between 0 and 1 and represent a portion of the nonzero area inside a certain neighborhood. Figure 6 shows the resulting FSS_{wind} value at different neighborhood sizes. As expected, the FSS_{wind} value tends to increase with neighborhood size and eventually reaches the asymptotic value of about 0.8, which indicates the existence of bias.

c. The behavior of the new score in selected examples

To better understand the behavior of the score, it is helpful to examine some specially selected examples.

From the MesoVICT dataset, a set of four interesting examples are selected for analysis. For each example in the set, the observational and the forecast wind class index fields are derived (Figs. 7a–d) and the resulting FSS_{wind} values calculated (Fig. 7e). The figures showing the original wind vector fields for these examples are available in the online supplemental material.

Example A represents an event with the smallest FSS_{wind} value at the smallest neighborhood ($n = 1$). Here convective events are mainly occurring along the stationary air mass boundary slightly to the north of the Alps. In the analysis, there is only a small number of convective cells, with winds in the surrounding areas generally being weak. In the model, the instability is larger over the majority of the studied region, leading to a higher number of convective cells that cause stronger winds. Because of the differences in VERA and forecasted winds fields, the same colored regions in Fig. 7a show very little overlap, resulting in a very low FSS_{wind} value of 0.28 (at $n = 1$). As the neighborhood

FIG. 5. An example of binary wind class fields that are used to calculate the fractions. The figure represents the same case as in Fig. 4.

increases, the FSS_{wind} value also increases and eventually reaches 0.82, reflecting a noticeable bias.

Example B represents the opposite of example A, an event with the largest FSS_{wind} value at the smallest neighborhood. Here two cold fronts are passing the area to the north of the Alps, with one going over the eastern Alps and initiating strong stationary thunderstorms in the area of Slovenia. The model forecasts the wind well for most of the domain, with the exception of limited areas to the south and east of the Alps where the model fails to forecast weak winds. In reality, stationary thunderstorms form in these regions, bringing local flooding. Overall, the same colored regions in VERA and the model show a good overlap over most of the domain. The good overlap results in a high FSS_{wind} value of 0.71, even at the smallest neighborhood. As the neighborhood increases, so does the FSS_{wind} value, which eventually reaches a near-perfect score value of 0.98, reflecting a very small bias.

Example C is chosen similarly to example A (an event with the smallest FSS_{wind} value at the smallest neighborhood), with the additional requirement that the asymptotic value is nearly perfect ($FSS_{wind}^{asy} > 0.98$). This extra requirement means the selected event will have no bias and, as a result, the FSS_{wind} value will only be influenced by the spatial displacement of the wind in the forecast. This is another instance of convection with a moist and unstable subtropical air mass causing convection during the day. Overall, the forecast is quite good with convection appearing in both the analysis and forecast, but locally the forecast is relatively poor since the convective cells' positions are hard to correctly predict with the model. The corresponding wind classes cover a very similar total area in VERA and the model (similar to example B) but, because of the significant spatial error, the areas are located in the wrong places. Thus, the FSS_{wind} value at small neighborhoods is poor (about 0.35) but, at the same time, the value improves toward a near-perfect value at

larger neighborhoods (about 0.99), reflecting an overall good forecast at large scales.

The last example, D, shows an event with the lowest asymptotic value. Here a situation with a squall line ahead of a cold front led to widespread convective events in large parts of the area with strong and gusty winds observed in the vicinity of the convective cells. Similarly to some previous examples, the model has problems forecasting the correct position of the convective cells. Moreover, because of the wrong timing of the passage of the weather front, the convection starts earlier in the model, resulting in a squall line and overall stronger winds with the analysis containing only a few isolated convective cells with weak winds in the surrounding regions. In the VERA analysis, yellow (calm/low wind) covers the majority of the domain. In the model, almost no yellow is present with green, red, and blue covering areas of comparable size. This discrepancy represents a significant bias of the forecast, which in turn causes a very low FSS_{wind}^{asy} value of 0.49. Nevertheless, some overlap does exist, most notably with the pink and blue regions near the northwest and

FIG. 6. The FSS_{wind} score value for the case shown in Figs. 4 and 5.

FIG. 7. Analysis of selected 1-h cases from the MesoVICT data. The cases are selected from having the best/worst FSS_{wind} value at the smallest neighborhood or asymptotic value. (a)–(d) Wind class index fields for the VERA analysis and the model forecast. (e) The FSS_{wind} value.

southeast corners, resulting in a low FSS_{wind} value of 0.3 (at $n = 1$).

d. The case-averaged score value

The FSS_{wind} value is calculated for each model and each 1-h time step and then averaged over all the time steps for a specific case to obtain the case-averaged FSS_{wind} value. The results are shown in Fig. 8. The case-averaged FSS_{wind} value shows a monotonic increase with neighborhood size for all cases and models. The values at the smallest neighborhood ($n = 1$) tend to be between 0.4 and 0.6, while the asymptotic values surpass the 0.9 value for all models and cases with the exception of the CMH_06 model for cases 3 and 4 (for which the value is around 0.85). The high asymptotic value indicates a small overall bias of wind classes. For cases 1, 3, and 4, the forecast made by CMH_06 clearly underperforms compared to the forecast by the CO models. The statistical significance of the difference in the models' performance is tested using a two-sample Kolmogorov–Smirnov nonparametric test (Wilks 2006). The test compares the distributions of the FSS_{wind} values of two

models. If the distributions are sufficiently different, the null hypothesis (that the samples are drawn from the same distribution) can be rejected. Since the FSS_{wind} value depends on the neighborhood size, the test is done separately for each neighborhood size. The user should note that the Kolmogorov–Smirnov test assumes the samples are independent. However, the fields and the resulting FSS_{wind} values will oftentimes be temporally correlated if consecutive fields are analyzed, which reduces the power of the test and care needs to be taken on how the statistical significance results are interpreted. In this analysis the fields are assumed to be uncorrelated although this might not be the case since the time difference between consecutive time steps is only one hour.

The test shows that, when comparing the CHM_06 with the other two models, the distributions of the FSS_{wind} values are statistically different (at $p = 0.01$) at all neighborhood sizes for cases 3 and 4, while for case 1 the differences are statistically significant for neighborhoods in the range from 7 to 85 grid points. The difference between CO2_00 and CO2_06 is not statistically significant for any case, indicating that the two models

FIG. 8. The case-averaged FSS_{wind} value for the six cases from the MesoVICT project.

perform with comparable skill. Case 2 is somewhat special since the CMH_06 performs the worst at small scales but is best at large scales (the differences are statistically significant), indicating a lower bias. For case 5, all models perform comparably and the differences are not statistically significant at any neighborhood size. For case 6, the models perform comparably at smaller neighborhoods whereas at larger neighborhoods the CMH_06 performs better because of the lower bias (the differences are statistically significant).

e. Analysis of sensitivity to class definitions

In the previous sections, the five basic wind classes defined in Fig. 3b are used. It is reasonable to assume the score value will depend on the chosen thresholds that define the classes, which is especially problematic for the wind direction thresholds. It makes sense to select the direction thresholds so that the direction-dependent classes span azimuthal ranges of equal size, although the exact location of the thresholds seems quite arbitrary. The selection of direction thresholds holds the potential to influence the score quite strongly. For example, there might be a case in which the prevailing wind direction would be very close to the threshold value used to distinguish between north and east wind. In this case, a small change in the wind direction in the forecast could dramatically change the FSS_{wind} value. Moreover, a relatively small error in the forecasted wind

could cause the score to have a very poor score value, indicating a very poor forecast, whereas in fact the error would be quite small. Such behavior would clearly not be in line with a subjective evaluation of the forecast made by a human analyst.

To test the sensitivity of the FSS_{wind} to azimuthal class rotation, an analysis of the MesoVICT dataset is performed in which an azimuthal rotation of the original five basic wind classes is implemented (Fig. 9a). For each rotation, the FSS_{wind} values are recalculated and compared to the values for other rotations. Figure 9b shows the sensitivity analysis for the example case shown in Figs. 4–6. The FSS_{wind} value is calculated for the original ($\gamma = 45^\circ$) class definition and then for every rotation between $\gamma = 0^\circ$ and $\gamma = 90^\circ$ (full azimuthal class width) in steps of 5° . The spread (the difference between the maximal and the minimal FSS_{wind} value) of the thin gray lines represent the spread of possible FSS_{wind} values. The spread is largest near the neighborhood size 10 (about 0.1) and is smallest at the largest neighborhoods (about 0.04). The maximal possible spread of 0.1 is relatively large since it represents 10% of the possible score value, demonstrating that a problem can indeed arise if only a single rotation of the class definition is used to calculate the score.

Luckily, this potentially problematic effect can be minimized by calculating the FSS_{wind} value a second time, but this time with the wind direction thresholds

FIG. 9. (a) Azimuthal rotation of the five basic wind classes. Angle $\gamma = 45^\circ$ represents the original rotation shown in Fig. 3b. (b) Sensitivity to the azimuthal rotation of class definition for the case shown in Figs. 4–6. The thin lines represent values for rotations in steps of $\Delta\gamma = 5^\circ$. The strong black lines represent rotations $\gamma = 0^\circ$ and $\gamma = 45^\circ$. (c) The percentile values of the difference between the maximum FSS_{wind} value and the minimum FSS_{wind} value are shown in black lines. The gray lines are related to the sensitivity of the approach, which uses the $\gamma = 0^\circ$ and $\gamma = 45^\circ$ rotations. First the values for the $\gamma = 0^\circ$ and $\gamma = 45^\circ$ rotations are calculated and the largest value of the two is selected. This value is then subtracted from the maximum FSS_{wind} value to obtain a difference. The gray lines represent the percentiles of this difference.

rotated by half the azimuthal class width (in this case calculated for $\gamma = 0^\circ$ and $\gamma = 45^\circ$). After the two FSS_{wind} values are calculated, only the larger value of the two can be used to avoid the potentially very poor score value that can result because of threshold sensitivity. This procedure brings the results more in line with a subjective evaluation of the forecast made by a human analyst. In Fig. 9b, the FSS_{wind} values for the $\gamma = 0^\circ$ and $\gamma = 45^\circ$ rotations are shown by thicker black lines. For this case, the $\gamma = 45^\circ$ variant represents the highest values, while a $\gamma = 0^\circ$ variant represents the lowest values. Since the $\gamma = 45^\circ$ variant has larger values at all neighborhoods, its values should be used to define the final score value.

This approach is tested over all 896 possible Mesos-VICT dataset comparisons between the analysis and the model. For each comparison and at each neighborhood

size, the difference between the maximal and minimal FSS_{wind} value is calculated (again using steps of 5°). This difference represents the largest possible magnitude of the sensitivity to class rotation. The percentiles of the largest possible sensitivity magnitude are shown in black in Fig. 9c. The sensitivity tends to be larger at the smaller neighborhoods with 1% of the comparisons (99th percentile) having a magnitude larger than 0.2 while the median magnitude (50th percentile) is about 0.06. These sensitivity magnitudes can be compared to the sensitivity when using the approach with the $\gamma = 0^\circ$ and $\gamma = 45^\circ$ rotations (shown in gray in Fig. 9c). In this case, 99% of the comparisons have a sensitivity magnitude smaller than 0.05 while the median sensitivity is smaller than 0.01. Since 0.01 represents only 1% of the possible score value, this approach can indeed be used to minimize the sensitivity of the score to the azimuthal rotation of class

FIG. 10. Alternative wind class definitions.

definitions. Nevertheless, there can be certain situations in which the sensitivity can be significantly larger. It is suggested that when the user suspects this behavior, the score values should be calculated for all rotations using a 5° step, and the largest value should be chosen. However, this situation greatly increases the computational cost of the score calculation.

In addition to the sensitivity to class rotation, there is another kind of sensitivity that is linked to the number of classes. In the case of Fig. 3b, five wind classes consist of a calm/low wind class and four additional classes according to the wind direction. Yet there is no inherent reason the classes have to be defined in this way. The classes could easily be made more sensitive to the wind direction and an additional four classes representing the northwest, northeast, southwest, and southeast directions included, with each class spanning only a 45° azimuth range (Fig. 10a). In this case, there would be a total of nine classes. Alternatively, additional classes could also be included according to wind speed, for example, 4 additional classes for “strong” winds as shown in Fig. 10b (e.g., the original easterly wind class is split into a “strong easterly wind” class and “medium easterly wind” class), or both modifications could be applied at the same time, resulting in 17 classes (Fig. 10c). When including additional classes, it is reasonable to expect that the overall FSS_{wind} value will tend to fall since the chance of an overlap will decrease simply because of the existence of more classes.

The impact of using alternative definitions for classes is tested on the MesoVICT cases. Figure 11 shows the analysis for cases 4 and 5, in which values for CMH_06 and CO_00 models are shown. Cases 4 and 5 are shown because they give very clear results when using the original class definition (the difference in model performance is either statistically significant or not, at all neighborhood sizes). When the number of classes increases, the FSS_{wind} value does indeed decrease. Thus, the score values for the original definition (5 classes) are the highest while the values for variant C (17 classes) are the lowest. The score values for variants A and B (both 9 classes) are between the values of the original and variant C with variant B having larger values than those of A at small neighborhoods and vice versa at larger neighborhoods. However, although the score values decrease when a higher number of classes are used, the conclusion is the same for all variants. Specifically, for case 4, the CO_00 model clearly performs better than the CMH_06 model (the difference in score value is statistically significant at all neighborhood sizes) while for case 5 there is no statistically significant difference between the models’ performance (at any neighborhood size).

FIG. 11. The impact of alternative wind class definitions on the FSS_{wind} value for the MesoVICT cases 4–5. The lines marked in black represent the CMH_06 model while the gray lines represent the CO2_00 model. The designation of class definitions with letters A–C corresponds to the class definitions shown in Fig. 10.

4. Discussion and conclusions

A new wind verification method is presented and analyzed. The new score is based on the idea of the fractions skill score. First, the original wind vector fields are decomposed into a number of binary fields using predefined wind classes. Second, the fractions are calculated, and the FSS_{wind} value is calculated similarly to the original FSS, but with the additional sum over all the classes.

Analysis of the MesoVICT cases and the idealized setup is performed to identify some properties of the new score. The new score avoids the problems of the non-spatial verification metrics (the double-penalty problem

and the failure to distinguish between a near miss and much poorer forecasts) and can distinguish the forecasts even when the spatial displacement of wind patterns is large. Moreover, the time-averaged FSS_{wind} value in combination with a statistical significance test (e.g., the two-sample Kolmogorov–Smirnov test) enables different wind forecasts to be ranked by their performance.

The FSS_{wind} value will always be influenced by the bias (if it exists) as well as the spatial error of the wind field in the forecast. The score value will be most influenced by the most frequent wind class present in the data while classes that appear only rarely will not significantly influence the score value. Since the score value will depend greatly on how the wind classes are defined, special care is needed when deciding on how the classes will be defined. Moreover, to alleviate the effect of sensitivity to class rotation the approach with the rotated classes should be used—this correction is shown to reduce the average magnitude of the sensitivity to about 1% of the score value.

It is worth noting that the new score can also be used with variables other than wind. Equations (1) and (2) are, in fact, general and represent a general multiclass version of the original FSS. The classes could also be defined for other variables. For example, instead of using a single precipitation class (as in the original FSS), two classes of precipitation could be used simultaneously: one representing medium-intensity precipitation and the other representing high-intensity precipitation. Alternatively, for atmospheric pressure, different classes could be defined, representing areas with low and high atmospheric pressure.

Acknowledgments. The authors wish to thank the three anonymous reviewers for a thorough examination of the manuscript and many valuable suggestions, which made the revised version of the manuscript much better than the original. The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding P1-0188).

REFERENCES

- American Meteorological Society, 2017: “Cardinal winds.” Glossary of Meteorology, http://glossary.ametsoc.org/wiki/Cardinal_winds.
- Brown, B. G., E. Gilleland, and E. E. Ebert, 2012: Forecasts of spatial fields. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed. I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 95–117, <https://doi.org/10.1002/9781119960003.ch6>.
- Dorninger, M., M. P. Mittermaier, E. Gilleland, E. E. Ebert, B. G. Brown, and L. J. Wilson, 2013: MesoVICT: Mesoscale Verification Inter-Comparison over Complex Terrain. NCAR Tech. Note NCAR/TN-505+STR, 23 pp., <https://doi.org/10.5065/D6416V21>.
- Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteor. Appl.*, **15**, 51–64, <https://doi.org/10.1002/met.25>.
- Faggian, N., B. Roux, P. Steinle, and B. Ebert, 2015: Fast calculation of the fractions skill score. *Mausam*, **66**, 457–466, http://metnet.imd.gov.in/mausamdocs/166310_F.pdf.
- Gilleland, E., 2013a: Testing competing precipitation forecasts accurately and efficiently: The spatial prediction comparison test. *Mon. Wea. Rev.*, **141**, 340–355, <https://doi.org/10.1175/MWR-D-12-00155.1>.
- , 2013b: Two-dimensional kernel smoothing: Using the R package “smoothie.” NCAR Tech. Note NCAR/TN-502+STR, 17 pp., <https://doi.org/10.5065/D61834G2>.
- , D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Wea. Forecasting*, **24**, 1416–1430, <https://doi.org/10.1175/2009WAF222269.1>.
- , —, —, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bull. Amer. Meteor. Soc.*, **91**, 1365–1376, <https://doi.org/10.1175/2010BAMS2819.1>.
- Harrison, G., 2014: *Meteorological Measurements and Instrumentation*. Wiley, 278 pp.
- Mittermaier, M., and N. Roberts, 2010: Intercomparison of spatial forecast verification methods: Identifying skillful spatial scales using the fractions skill score. *Wea. Forecasting*, **25**, 343–354, <https://doi.org/10.1175/2009WAF222260.1>.
- , —, and S. A. Thompson, 2013: A long-term assessment of precipitation forecast skill using the fractions skill score. *Meteor. Appl.*, **20**, 176–186, <https://doi.org/10.1002/met.296>.
- Roberts, N., 2008: Assessing the spatial and temporal variation in the skill of precipitation forecasts from an NWP model. *Meteor. Appl.*, **15**, 163–169, <https://doi.org/10.1002/met.57>.
- , and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, <https://doi.org/10.1175/2007MWR2123.1>.
- Rossa, A., P. Nurmi, and E. E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. Michaelides, Ed., Springer, 419–452, https://doi.org/10.1007/978-3-540-77655-0_16.
- Skok, G., 2015: Analysis of fraction skill score properties for a displaced rainband in a rectangular domain. *Meteor. Appl.*, **22**, 477–484, <https://doi.org/10.1002/met.1478>.
- , 2016: Analysis of fraction skill score properties for a displaced rainy grid point in a rectangular domain. *Atmos. Res.*, **169**, 556–565, <https://doi.org/10.1016/j.atmosres.2015.04.012>.
- , and N. Roberts, 2016: Analysis of fractions skill score properties for random precipitation fields and ECMWF forecasts. *Quart. J. Roy. Meteor. Soc.*, **142**, 2599–2610, <https://doi.org/10.1002/qj.2849>.
- Steinacker, R., C. Häberli, and W. Pötschacher, 2000: A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data. *Mon. Wea. Rev.*, **128**, 2303–2316, [https://doi.org/10.1175/1520-0493\(2000\)128<2303:ATMFTA>2.0.CO;2](https://doi.org/10.1175/1520-0493(2000)128<2303:ATMFTA>2.0.CO;2).
- Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.