Robust Worst-Case Scenarios from Ensemble Forecasts

SEBASTIAN SCHER,^{a,b,c} STEPHEN JEWSON,^d AND GABRIELE MESSORI^{e,f,a,b}

^a *Department of Meteorology (MISU), Stockholm University, Stockholm, Sweden*

^b *Bolin Centre for Climate Change Research, Stockholm University, Stockholm, Sweden*

^c *Know-Center GmbH, Graz, Austria*

^d *London, United Kingdom*

^e *Department of Earth Sciences, Uppsala University, Uppsala, Sweden*

^f *Centre of Natural Hazards and Disaster Science (CNDS), Uppsala University, Uppsala, Sweden*

(Manuscript received 26 November 2020, in final form 26 April 2021)

ABSTRACT: To extract the most information from an ensemble forecast, users would need to consider the possible impacts of every member in the ensemble. However, not all users have the resources to do this. Many may opt to consider only the ensemble mean and possibly some measure of spread around the mean. This provides little information about potential worst-case scenarios. We explore different methods to extract worst-case scenarios from an ensemble forecast, for a given definition of severity of impact: taking the worst member of the ensemble, calculating the mean of the N worst members, and two methods that use a statistical tool known as directional component analysis (DCA). We assess the advantages and disadvantages of the four methods in terms of whether they produce spatial worst-case scenarios that are not overly sensitive to the finite size and randomness of the ensemble or small changes in the chosen geographical domain. The methods are tested on synthetic data and on temperature forecasts from ECMWF. The mean of the N worst members is more robust than the worst member, while the DCA-based patterns are more robust than either. Furthermore, if the ensemble variability is well described by the covariance matrix, the DCA patterns have the statistical property that they are just as severe as those from the other two methods, but more likely. We conclude that the DCA approach is a tool that could be routinely applied to extract worst-case scenarios from ensemble forecasts.

KEYWORDS: Ensembles; Operational forecasting; Probability forecasts/models/distribution; Decision support

1. Introduction

In many weather forecasting settings forecast users are interested in the most likely weather at a specific lead time, potentially with an added measure of uncertainty. A commonly used tool providing this information is ensemble forecasting (e.g., [Leutbecher and Palmer 2008](#)). In an ensemble forecast, a set of individual forecasts is produced (typically 10–50 forecasts). The individual forecasts either have slightly different initial conditions, and/or slightly different model formulations or stochastic perturbations. The ensemble of forecasts can then be analyzed statistically. For example, the ensemble mean is frequently used as the best-guess forecast of the future weather state. Ensemble forecasts have been in use for over two decades in many different fields, including for global medium-range weather forecasts ([Molteni et al. 1996](#)), short-term temperature and air pollution forecasts ([Stensrud and Yussouf 2003](#)), wind power production ([Taylor et al. 2009](#)), flood forecasting ([Rossa et al. 2011](#)), regional high-resolution precipitation forecasts ([Vié et al. 2012](#)), ship routing ([Hoffschmidt et al. 1999](#); [Chu et al. 2015](#)), and forecasting icing on wind turbines ([Molinder et al. 2018](#)). Ensemble forecasts

are used by a range of user groups, not limited to experts in weather forecasting ([Fundel et al. 2019](#)), and their economic benefit compared to single-value forecasts (i.e., single best-guess forecasts) has been clearly shown (e.g., [Richardson 2000](#); [Palmer 2002](#)).

A thorough approach to considering the uncertainty in the impacts associated with an ensemble forecast would be to look at the possible impacts of every ensemble member. However, in many cases, and especially when considering forecasts that consist of spatial patterns, this may not be feasible or practical. Converting even a single ensemble member to an estimate of the resulting impacts may be a costly process if the impact calculation is complex or not fully automated. In such cases, users may be limited to looking at the impact of just the ensemble mean. However, using only the ensemble mean risks missing key information regarding worst-case spatial scenarios. This is unfortunate, since for many decision problems it is the low-probability but high-impact worst-case events in the forecast that are most important ([Palmer 2002](#)). For example, during the summer months a national health authority may want to know “How hot might it get next week across the country?” yet have no routine in-house way of extracting this information from an ensemble forecast. Similarly, emergency services or civil protection agencies may want to prepare in advance for worst-case weather scenarios, whether heatwaves, cold spells, heavy precipitation, or other. In such cases, an approach that avoids the complexity of using every ensemble member but nevertheless provides some information

Denotes content that is immediately available upon publication as open access.

Corresponding author: Sebastian Scher, sscher@know-center.at

DOI: 10.1175/WAF-D-20-0219.1

© 2021 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy](#) (www.ametsoc.org/PUBSReuseLicenses).

about possible extreme spatial patterns may be valuable. Motivated by this idea, we investigate methods for identifying a single pattern that reflects a representative worst-case deviation from the ensemble mean that could be considered in addition to the ensemble mean.

To discuss how to derive such a worst-case scenario, we first need to define “worst,” for which we need to define “impact.” For the sake of a general discussion and straightforward presentation of the ideas we just use a simple example based on temperature. We represent the notion of impact using average temperature over the domain, with the assumption that higher values of domain-averaged temperature have a higher impact. This simple example is sufficient for illustrating and testing different worst-case methodologies, and does have some real-world relevance, as heat-waves can have a strong impact on mortality (e.g., Fouillet et al. 2006). For specific applications, different measures of impact would be appropriate. For instance, actual temperatures could be replaced by temperature anomalies relative to a climatology, which would be a more appropriate measure of impact in some cases. Also, temperature could be replaced by other variables, or functions and combinations of variables. We discuss the applicability of the approaches we investigate to different types of forecast variables in detail in section 2a below. In addition to using different variables, the use of the average value over the domain could also be replaced by a weighted average, where the weights could be derived from exposure variations in space, or from linearization of a nonlinear impact function.

We also need to establish some terminology. We will use the word “plausible” to describe the extent to which an ensemble member or spatial pattern is likely or not. We use the term “extreme” to refer to extremes of the forecast distribution, or in other words values which are extreme within the ensemble, relative to the other ensemble members. Depending on the forecast, these may or may not be extremes relative to the unconditional distribution for the variable for that time of year. We use the term “robust” to indicate low sensitivity of the results to factors such as the errors induced by the fact that we have only a finite number of ensemble members, the errors introduced by only using a subset of those ensemble members to calculate the worst-case, and the sensitivity to small changes in the geographical domain.

Given our chosen definition of impact, we investigate several approaches for deriving worst-cases for forecasts consisting of spatial patterns. If one only cares about forecasts for each individual location, the most obvious way to extract information about extremes from an ensemble forecast is to compute local empirical percentiles of the ensemble (e.g., the 5th or 95th percentile). Alternatively, one can fit distributions to the ensemble at each location and then compute the percentiles of these distributions. However, for creating realistic spatial patterns of possible future weather, simply combining percentiles from each location to make a spatial pattern is not appropriate since the resulting pattern would typically be unlikely to occur in reality, and may even be impossible. This is because the variability within the ensemble across the different forecast locations is unlikely to be highly positively correlated for any except the smallest domains. Other methods that deal

with the problem of interpreting ensemble forecasts, extracting reasonable forecast scenarios and identifying extreme occurrences have been described by previous authors. One is a clustering method used among others by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Ferranti and Corti 2011; ECMWF 2015). In this method ensemble forecasts for a region are grouped by a clustering algorithm and each cluster is represented by the member that is closest to the centroid of that cluster. However, this method does not focus on the extreme cases of the ensemble, nor is there any categorization of the clusters as being more or less extreme, since there is no guarantee that there is a cluster that contains all extreme members. As a result, this method serves a different purpose from the methods tested in our study. Another method, that specifically focuses on identifying extremes in forecasts, is the extreme forecast index (EFI; Lalaurette 2003), also used by ECMWF. EFI is a measure of how extreme the ensemble forecast distribution is compared to the model climate. It is, however, computed on a point-by-point basis, and is not designed to deliver plausible spatial patterns. From the point of view of creating plausible patterns it therefore has the same shortcomings as locally computed percentiles. We therefore do not consider local percentiles or EFI as useful method for deriving realistic worst-case spatial patterns and focus on techniques that attempt to create more realistic patterns in space. We present and test four such techniques.

The first of the four methods we consider simply involves taking the worst ensemble member. Given our definition of impact, the worst member is the ensemble member with the highest average temperature across the domain. The second method involves taking the mean of the N worst members. The third and fourth methods are a novel application of directional component analysis (DCA) (Jewson 2020) that, to the best of our knowledge, has not been used for this purpose before. DCA provides a spatial pattern that maximizes likelihood for a given linear impact function. It provides no intrinsic magnitude information and hence the spatial pattern may be scaled by an arbitrary factor. We consider two different scaling factors, which we discuss in detail in section 2. We test our methods both on synthetic data and on operational temperature forecasts from ECMWF

2. Methods

In this section we describe the data we use, the methods for estimating worst-case scenarios, and how we evaluate the robustness of these scenarios.

a. Data

1) SYNTHETIC DATA

The first part of our analysis involves using simple simulations to develop our intuition as to the likely behavior of the four worst-case scenario methods we are investigating. We do this by simulating synthetic two-dimensional forecasts, to be interpreted as one forecast made for two locations simultaneously. In the first set of simulated values we produce the forecast values at the two locations are simulated from

uncorrelated normal distributions with mean 0 and standard deviation 1. We create ensembles of these forecasts, with each simulated ensemble containing 50 members, and in total we simulate 10 000 ensembles. These represent 10 000 realizations of a 50-member ensemble for the same forecast situation, and allow us to understand how an ensemble, and diagnostics derived from an ensemble, can vary simply due to the randomness in the ensemble-generating process. We repeated these analyses using correlated normal distributions and the results were qualitatively very similar (not shown).

To evaluate whether the results derived using normally distributed data are sensitive to the choice of distribution, we generate two additional sets of simulated ensembles. The first uses independent gamma distributions for the two locations, each with shape parameter of 2 and scale parameter of 1. This is a distribution with a large positive skew. The second set uses uniform distributions at the two locations, simulated such that in any single ensemble member one location is zero and the other location is nonzero. This is to simulate what might be seen in an ensemble of predictions of convective rainfall at two locations that are sufficiently far apart that individual convective events cannot affect both locations.

2) FORECAST DATA

We use 2-m forecasts of instantaneous temperature from the operational ensemble model of ECMWF. We use forecasts initialized at 1200 UTC every fifth day from 1 June 2019 to 25 August 2019 (a total of 18 cases). The number of forecasts we use is sufficient to generate statistically significant results for most parts of our analysis (where we assess significance at the 95% confidence level). We focus on a lead time of 72 h. We repeated the analysis with a lead time of 48 h, and the qualitative results did not change (not shown). The domain is 5°W–30°E, 40°–60°N, which covers a large part of Europe. We do not apply any bias correction or forecast calibration. While uncalibrated ensemble forecasts usually have biases in their probabilities, this should not affect our results. We are simply assessing how, given an ensemble, one can extract plausible worst-case scenarios from the ensemble members. All of our four methods could equally well be used on an ensemble that has been calibrated prior to analysis.

b. Methods for worst-case scenario estimation

We now describe the four methods we use for creating worst-case scenarios in more detail.

1) WORST MEMBER

The first approach we consider is to use the worst ensemble member as the worst-case scenario. We label this method W1. Given our definition of impact as average temperature, the worst member is that with the highest domain-averaged temperature. Using the worst ensemble member as the worst-case scenario requires no assumptions about the statistical behavior of the ensemble and is easy to interpret, since there is no averaging of members, and we are simply taking one realization of the forecast model. We expect this method not to be very robust, however, as it is based on a single member of the ensemble and ignores information from the other members. If

one were to rerun the ensemble (with different initial random seeds, or a different number of members) the worst member could change significantly. It also does not converge as the size of the ensemble increases.

2) MEAN OF N WORST MEMBERS

The second approach we consider is to take the N worst members and average them. We use $N = 5$ (out of 50 members) and label this method W5. This should be more robust than considering a single member. If the size of the ensemble is increased, this method will converge, provided the value of N is adjusted accordingly to represent a constant proportion of the total ensemble size. In general, this method is only valid if it is reasonable to make the assumption that averaging together ensemble members creates patterns which are also plausible as ensemble members. If this assumption does not hold, then averaging ensemble members may create an unphysical pattern that could never be produced by the forecast model.

The W1 and W5 methods are illustrated in the sketch in Fig. 1a for a hypothetical forecast along several points of longitude.

3) DIRECTIONAL COMPONENT ANALYSIS

The third and fourth approaches we consider use directional component analysis (DCA). DCA—not to be confused with discriminant component analysis—is a method that has been proposed as a complement to principal component analysis (PCA) to find patterns of extremes (Jewson 2020). Both PCA and DCA are calculated from the covariance matrix of the data, where in our case the data are the ensemble forecast. While PCA finds patterns that maximize a certain metric of variability, DCA finds patterns that maximize the sum of the spatial field for a given level of likelihood, and is hence potentially more appropriate for finding extreme spatial patterns. The optimality properties of PCA and DCA hold true when the variability in the data to which they are applied is well-captured by the covariance matrix. In statistical terms, this means that the multivariate distribution of the data is well modeled as an elliptical distribution, such as a multivariate normal or multivariate t distribution.

The first DCA pattern is computed as follows. Given an ensemble forecast $x_{i,j}$ at grid points $i = 1, \dots, L$ (where latitude and longitude are flattened into a single dimension with length L), with members $j = 1, \dots, M$, we compute anomalies $\mathbf{a}$ with respect to the ensemble mean as follows:

$$a_{i,j} = x_{i,j} - \bar{x}_i. \quad (1)$$

We then compute the spatial covariance matrix $\mathbf{C}$:

$$C_{i,k} = \frac{1}{M} \sum_{j=1}^M a_{i,j} a_{j,k}^T = \frac{1}{M} \sum_{j=1}^M a_{i,j} a_{k,j}, \quad (2)$$

and the first DCA vector $\mathbf{g}$ is defined as follows:

$$\mathbf{g}_i = \sum_{k=1}^L C_{i,k}. \quad (3)$$

FIG. 1. (a) Illustration of worst member (W1; continuous orange curve) and mean of the five worst members (W5; dotted green curve) for a hypothetical 1D temperature forecast. (b) Illustration of DCA for a two-variable problem. The dotted arrow shows the direction of the (first) DCA pattern; the green and orange dots show W5 and W1, respectively; the diagonal lines show lines of constant severity of impact, with higher impact to the top-right-hand corner; the length of the DCA arrows have been set so that they match the severity of W5 and W1; the ellipse shows a line of constant probability density, with larger values inside the ellipse; the double-lined arrow shows the direction of the first PCA pattern (not used in this study). (c) Definition of the angle α , which is a measure of the similarity between the (1, 1) vector and the vector pointing to the green dot, and is hence a measure of the similarity between a pattern of uniform temperature and the pattern corresponding to the green dot. (d) Definition of the impact a of the pattern given by the green dot, which is the mean value of the field (in this case the mean value of a_1 and a_2).

Given the above expressions, we can also write the first DCA pattern as follows:

$$g_i = \frac{1}{M} \sum_{k=1}^L \sum_{j=1}^M a_{i,j} a_{k,j} = \frac{1}{M} \sum_{j=1}^M a_{i,j} \sum_{k=1}^L a_{k,j} = \frac{1}{M} \sum_{j=1}^M a_{i,j} w_j, \quad (4)$$

where

$$w_j = \sum_{k=1}^L a_{k,j}. \quad (5)$$

Subsequent DCA patterns, orthogonal to the first pattern, can also be calculated. However, in this application we will only consider the first pattern, which we will simply refer to as the DCA pattern. From Eq. (3), we see that the DCA pattern is calculated by summing columns in the covariance matrix. From Eq. (4), we see that DCA can also be calculated as a weighted sum of all the M patterns in the ensemble, where the weights are proportional to the severity of each ensemble member, and are given by Eq. (5). This second expression does not involve

the covariance matrix, and hence blurs the question of whether this method is parametric or nonparametric. For the derivation and other details of DCA, see Jewson (2020). Note that Jewson computes the DCA over the time dimension, whereas we compute it over the ensemble member dimension. The principle of DCA is illustrated in Fig. 1b, for a two-variable example. This could for example be a forecast for two longitude locations from the illustration in Fig. 1a.

As with the W1 and W5 methods, DCA can be generalized to alternative definitions of impact. In the case where the measure of impact is a weighted average over different locations, DCA becomes a weighted sum over columns in the covariance matrix. DCA also generalizes to impact measured as a quadratic function of the underlying variable.

Scaling

The amplitude of the DCA pattern can be scaled in different ways to suit different purposes. To obtain a meaningful amplitude for a worst case, we rescale the pattern in two different ways, which creates the third and fourth methods that we test:

1) we scale it so that the area-mean anomaly of the DCA pattern equals the area-mean anomaly of the worst member, i.e., has the same scaling as the W1 pattern. We label this method DCA1; 2) we scale it so that the area-mean anomaly of the DCA pattern equals the area-mean anomaly of the mean of the five worst members, i.e., has the same scaling as the W5 pattern. We label this method DCA5. The reason for these choices of scaling is to create two different DCA worst-case scenarios, one comparable in severity to the W1 pattern and one comparable to the W5 pattern. By taking the scalings for DCA from the W1 and W5 methods we can focus on the differences in the patterns generated by the different methods rather than the scaling.

4) LOCAL PERCENTILES

A simple way to get a map of extremes is to compute a percentile at each grid point. Here we use the 95th percentile. While this makes sense for single location forecasts, for our purpose it is problematic when looking at a large geographical domain. Experiencing extreme weather everywhere simultaneously at continental scale is extremely unlikely, and hence a map of local percentiles gives an unrealistically “bad” (in our case, warm) worst-case scenario (see also the discussion in the introduction). This is therefore only presented as a reference method for comparison purposes, and we do not consider it as a method for generating genuine candidates for worst-case scenarios.

c. Estimating robustness

We perform two kinds of robustness estimation on the results from our four worst-case estimation methods. The first estimates the robustness to the finite size and randomness of the ensemble, while the second estimates the robustness to the exact choice of spatial domain.

1) STATISTICAL ROBUSTNESS

The first kind of robustness estimation we perform investigates the extent to which our four methods for generating worst-case scenarios are robust to the finite size and randomness inherent in the ensemble. We call this statistical robustness. For the synthetic data, this can be estimated by simply creating a large number of realizations of the ensemble for the same underlying distribution. However, estimating statistical robustness for the real forecast data is more difficult. The ideal procedure would be to rerun the ensemble forecast model several times, and quantify how the results change from run to run, analogously to the process used for the synthetic data. Since this is unfeasible, we use three alternative procedures (and we will call them *procedures* to differentiate them from the *methods* we are using to create the candidates for worst-case scenario). Each robustness estimation procedure has advantages and disadvantages, and none can be said to be globally better than the others. We will consider the results from all three robustness estimation procedures in parallel. All three procedures artificially generate new ensembles from the given ensemble, without actually running the forecast model again, and our evaluation of robustness is based on assessing how much the estimates of worst case vary when applied to these new ensembles. The three procedures work as follows:

- **Bootstrapping:** we randomly pick 50 out of 50 members from the original ensemble with the possibility of drawing the same member multiple times (i.e., sampling with replacement). The main limitation of this procedure is that the set of patterns stays the same. This is a problem for estimating the sensitivity of the W1 method, since we either pick the worst member of the full ensemble in our new artificial ensemble or we do not. In the cases where we do pick it, which will be most cases, the results for the W1 method do not change. As a result, this procedure will tend to overestimate the robustness of the W1 method.
- **Reduced ensemble bootstrapping (subensemble, or subens):** we randomly pick 25 out of 50 members from the original ensemble, this time without replacement. This procedure suffers from the same limitation as bootstrapping, in that the worst member may be included in the subensemble, in which case results for the W1 method will not change. However, because the probability of including the most extreme member is now reduced, this is less of a problem than it is for bootstrapping.
- **Multivariate normal distribution (MVN):** we fit a multivariate normal distribution to the ensemble, and then randomly draw 50 members from this distribution. This procedure avoids the limitations of bootstrapping and generates entirely new patterns. However, it only does so by making statistical assumptions about the distribution of the data which may not be fully accurate. Bootstrapping and reduced ensemble bootstrapping, on the contrary, do not make any assumptions about the distribution of the data.

With each procedure, 50 new artificial ensembles are created, and the worst-case analysis is repeated on each of these 50 ensembles. For each artificial ensemble and each worst-case method, we compute the mean anomaly a with respect to the ensemble mean over the whole domain, and the angle α of the anomaly vector with respect to the unit vector:

$$\alpha = \frac{\arccos(\mathbf{a} \cdot \mathbf{1})}{\|\mathbf{a}\| \|\mathbf{1}\|}. \quad (6)$$

These two methods are illustrated in a two dimensional case in Fig. 1. Changes in the mean a can be understood as changes in the overall severity, or impact, of the pattern. Changes in the angle α are a simple way of measuring the size of changes in the spatial pattern. A limitation of using α to measure changes in the spatial patterns is that it is possible that the spatial pattern can change while α remains the same. However, in most cases if the spatial pattern changes α would change and in general α provides a simple way of measuring the size of pattern changes with a single number. The amplitude a and the angle α are complementary to each other, in that α is independent of scaling. If the whole pattern is multiplied by a constant, then α does not change.

The difference between each possible pair of worst-case methods in terms of the mean of the standard deviation of a and α over all forecasts is tested with the one-sample Student's t test, with the null hypothesis that the difference is zero [for details on the t test, see for example Wasserman (2004) and Wilks (2011)].

FIG. 2. Results for synthetic data following normal distributions. (a),(b) Examples of 50-member ensemble forecasts (black), all from the same distribution, and the corresponding worst member (W1; orange), mean of the five worst members (DCA5; green), DCA scaled on worst member (DCA1, blue), and DCA scaled on mean of the five worst members (DCA5; pink). (c) W1 (orange) and DCA1 (blue) for 10 000 different ensembles from the same distribution. The blue dots are plotted on top of the orange dots. (d) W5 (green) and DCA5 (pink) for the same ensembles as in (c). The pink dots are plotted on top of the green dots.

2) ROBUSTNESS TO CHANGES IN THE DOMAIN

The second kind of robustness estimation we consider is robustness to small changes in the geographical domain over which the worst-case scenarios are calculated. For this, we vary the domain by adding and subtracting 2° from each side of the standard domain, in all possible combinations (resulting in 81 alternative domains). For each domain, α and α are computed.

3. Results

We now present results from applying our four worst-case scenario generation methods to the synthetic and real forecast data. We then present results from applying the robustness procedures to the real forecast data.

a. Synthetic data

The synthetic data allows us to generate a large number of ensembles for the same underlying forecast, from which we can assess the random variability in the ensemble and in quantities derived from the ensemble. Two examples of the ensembles we generate using normal distributions are shown in Figs. 2a and 2b, which show all the ensemble members (small dots in black) and the four different methods for defining worst-case scenarios (large dots in colors). There is no difference between the data in these two figures, except for different random seeds in the underlying calculation. That the DCA1 and DCA5 patterns are scaled to the same mean anomaly as the W1 and W5 patterns, respectively, is apparent in both panels. Figure 2c shows results from two of the worst-case methods (W1 and DCA1) from

FIG. 3. Example forecast 1 (forecast initialized at 1200 UTC 1 Jun 2019 with a lead time of 72 h). Shown are the worst-case scenarios, plotted as 2-m temperature anomalies (K) with respect to the ensemble mean, for the four methods plus local percentiles. Panel titles indicate the domain-averaged temperature anomalies.

each of the 10 000 simulated ensembles. As expected, the W1 approach is not very robust and shows large variability (orange dots in Fig. 2c). The DCA1 pattern shows the same variability in amplitude, as it must because it uses the same amplitude scaling, but much smaller variability in direction from the origin (blue dots in Fig. 2c, which are plotted on top of the orange dots). In the two-dimensional setup of the synthetic data, direction from the origin represents the shape of the spatial pattern (analogous to α in two dimensions) and so this smaller variability in the direction from the origin represents smaller variability in the spatial pattern. Figure 2d shows results from the other two worst-case methods (W5, green dots, and DCA5, pink dots plotted on top of the green dots), from each of the 10 000 simulated ensembles. The W5 pattern shows less variability in both amplitude and direction from the origin than the W1 results shown in Fig. 2c. The DCA5 pattern shows the same variability in amplitude as the W5 pattern, as it must, but smaller variability in direction from the origin i.e., smaller variability in spatial pattern.

The reduced variability in terms of direction from the origin of the DCA patterns relative to the W1 and W5 methods is a

strength of the DCA method, as it indicates that the spatial patterns would change less if the ensemble were run again, compared to the other methods. In this sense the DCA methods are more robust. The robustness of DCA arises because DCA is estimated from the covariance matrix of the ensemble, which in turn is based on all members of the ensemble. This means it is based on more data than W1 or the W5, which are based only on subsets of the ensemble. The differences between W1 and DCA are larger in Fig. 2c than the differences between W5 and DCA in Fig. 2d. This makes sense, as DCA uses 50 members to compute the spatial pattern, W5 uses five members to compute the spatial pattern, and W1 is based only on a single pattern. As a result, DCA uses 50 times as much data as W1, but only 10 times as much as W5.

The synthetic data used to create these results was generated using normal distributions, which is the most favorable situation for applying DCA. To test whether similar results hold for other distributions, we have repeated the analysis using the synthetic gamma-distributed data and the synthetic uniformly distributed datasets (see section 2). The gamma results, corresponding to the results shown in Fig. 2, are shown in Fig. A1

FIG. 4. As in Fig. 3, but for example forecast 2 (forecast initialized at 1200 UTC 21 Jul 2019, with a lead time of 72 h).

in appendix A. The results are very similar to the results for the normally distributed data shown in Fig. 2. W5 is more robust than W1, and DCA is more robust than either W5 or W1. The results for the uniformly distributed data are shown in Fig. A2 and are now very different. These data are simulated in such a way that no event can affect both locations—in analogy to the fact that convective rainfall will seldom occur simultaneously at two distant locations—and as a result all events lie on the horizontal and vertical axes. The DCA patterns, and W5, all involve averaging together events, which leads to patterns that have rain at both locations and are therefore not realistic. The ensemble mean, shown by an X, is also not a realistic pattern. W1 is a realistic pattern, but is only representative of one location. The critical difference between the gamma and uniform examples is not the distributions themselves, but the different assumptions about the dependencies between the variability at different locations, which leads to different behavior in terms of whether or not averaging together members of the ensemble creates realistic patterns.

We conclude that, even if the data are not normally distributed, but are such that averaging together ensemble members creates patterns that are plausible alternative ensemble members, then W5 and DCA remain reasonable, and

robust. If on the other hand the data are such that averaging together ensemble members creates patterns that are not realistic, then neither W5 nor DCA are appropriate as worst-case patterns (and even the ensemble mean is not a realistic pattern).

b. Real forecasts

Figures 3 and 4 show examples of applying all four worst-case methods to two individual ECMWF forecasts. We also show maps of local 95th percentiles. All panels show the anomalies with respect to the ensemble mean. For completeness, the verifying anomaly (observed anomaly with respect to the ensemble mean, from the ERA5 reanalysis Hersbach et al. 2020) is shown in Fig. B1 (see appendix B). It should, however, be pointed out that comparing the worst-case scenarios to the verification in individual cases can be misleading, as the worst-case scenarios have by definition low likelihood and thus only in rare cases would they be expected to be similar to the actual observed pattern.

In the first ECMWF forecast example, shown in Fig. 3, all four candidates for worst-case scenarios (i.e., excluding the map of percentiles), show patterns with broadly similar large-scale spatial features. The DCA1 pattern (Fig. 3c) is much

FIG. 5. Sensitivity of the worst-case scenario methods to the underlying ensemble for a single forecast (initialized at 1200 UTC 21 Jul 2019), estimated with three different robustness estimation procedures. (a) Bootstrapping, (b) subensemble, and (c) multivariate normal. The cross markers show the mean of each method; α has units of kelvins (K).

smoother than the W1 pattern (Fig. 3a). The DCA5 pattern (Fig. 3d) is somewhat smoother than the W5 pattern (Fig. 3b). The W1 pattern contains some small-scale anomalies over the United Kingdom and northern France that are entirely absent

in DCA1, and mostly absent in W5. W1 also contains large-scale positive and negative temperature anomalies in the eastern part of the domain that have weaker magnitudes in the other patterns.

FIG. 6. Sensitivity of the worst-case scenario approaches to the underlying ensemble over all forecasts. Panels show boxplots of standard deviation of amplitude σ and standard deviation of angle α for all forecast cases, for the three different robustness estimation procedures

FIG. 7. Sensitivity of the worst-case scenario approaches to changes in geographical domain, over all forecasts. Panels show boxplots of standard deviation of amplitude a and standard deviation of angle α for all forecast cases

In the second ECMWF forecast example, shown in Fig. 4, there is less consistency between W1 and the other patterns. For example, W1 (Fig. 4a) has a region of negative anomalies in the western Balkans, while W5 (Fig. 4b) and the DCA patterns (Figs. 4c,d) show positive anomalies there. The DCA1 pattern is again much smoother than W1, and the DCA5 pattern is slightly smoother than W5.

In both of the above forecast examples, we suspect that the small-scale features in W1 are unlikely to be robust features of the forecast, while the main features in DCA1 are likely to be more robust. This interpretation is based on the spatial scales involved, the lead time of the forecast (72 h), and the results from the synthetic data which showed much greater pattern variability in W1. This is an argument to prefer DCA1 over W1, as a representative example of a worst-case scenario. In addition, we know from the derivation of DCA that, if the statistical assumptions are correct, DCA1 will be more likely than W1. This is because the two patterns have the same amplitude, and DCA is by definition the pattern with the highest likelihood among all possible patterns of the same amplitude. The differences between the W5 pattern and the DCA5 pattern are similar, but smaller. Once again, we know from the derivation of DCA that the DCA pattern will be more likely, if the statistical assumptions are satisfied. The patterns constructed from local percentiles in both Figs. 3e and 4e show strong positive anomalies almost everywhere, which are extremely unlikely, if not impossible, in reality.

As a final observation, we note that none of the methods have large contributions from ocean regions. Ocean data points are not treated differently from land points in the analysis: their lack of contribution is due to the low variability of 2-m temperature over the ocean across the ensemble, likely because low-level air temperatures remain in reasonable equilibrium with sea surface temperatures in most cases. The low variability of the temperatures over the ocean is also

evident in the maps of the 95th percentiles (Figs. 3e and 4e). If there were a reason to focus on temperature variability over the ocean this could be achieved by using standardized anomalies or spatial weighting.

The ensemble we are testing contains 50 members, but many operational ensembles contain many fewer members. To assess the impact of fewer members on the results from our four worst-case methods, we repeat the above analysis on subsets of 10 members. The results are shown in Figs. C1 and C2 in appendix C. The comparison between W1 and DCA1 shows essentially the same behavior as for the results from the 50-member ensemble i.e., the W1 pattern shows small-scale variability that may not be robust and the DCA1 pattern is much smoother. The W5 pattern is still defined as the worst five members, which is half the ensemble for a 10-member ensemble. The W5 pattern and the DCA5 pattern look very similar, and more similar than they do for the 50-member ensemble, because they are now defined in a more similar way.

c. Statistical robustness

We now look at the robustness of the results to the randomness and finite size of the ensemble, using the procedures described in section 2 above. Figure 5 shows the sensitivity of amplitude (or impact) a and angle α of a single forecast, for the 3 different robustness estimation procedures. Each point shows the amplitude and angle for a single new ensemble generated from the original ensemble by one of the procedures. Since these are results for a single forecast, they cannot tell us about the general statistical behavior of the different worst-case scenario methods. However, we can see indications of behaviors that we would expect, based on the results from the synthetic data. For instance, the results from W1 and DCA1 (blue dots) show greater variability in amplitude than the results from W5 and DCA5 (green dots), for all three procedures.

FIG. A1. As in Fig. 2, but for synthetic data whose two marginal distributions are gamma distributions with shape parameter of 2 and scale parameter of 1. This gives distributions that are strongly positively skewed. The data are then standardized, for comparison with Fig. 2. The axis ranges are increased, to allow for the wider range of maximum values that occurs with the gamma distribution.

Also, the variability of the angle of the DCA methods is lower than the variability of the other two methods.

Figure 6 is based on the same data as Fig. 5, but collected for all 18 forecast cases that we consider, in order to understand the general statistical behavior of the different worst-case scenario methods. For each forecast, the standard deviations of severity a and the angle α are estimated, and the figure shows boxplots of these standard deviations. High standard deviation means high sensitivity and low robustness, and low standard deviation means low sensitivity and high robustness. For the amplitude (Fig. 6a) W5 has a lower sensitivity than W1 for all three procedures for estimating the sensitivity, and statistically significantly so (see Table D1 in appendix D, which shows the p values for the difference of the mean standard deviation), small p values indicate that the two are significantly different at a high confidence level). The DCA patterns have—by

design—the same amplitude sensitivity as the method they are scaled on. For the angle (Fig. 6b), we see that the two DCA patterns have the same value of standard deviation within each procedure, as the angle is independent of scaling. The sensitivity of the angle of the DCA patterns is lower than that of both W1 and W5, for all procedures for estimating sensitivity. Compared to W5, the difference is significant (see Table D2, which shows the p values for the difference of mean standard deviations). Whether the lower sensitivity of the pattern of DCA compared to W1 is also statistically significant depends on the sensitivity estimation procedure. When using the MVN-uncertainty procedure the differences are significant, but with the other two procedures they are not. For the difference between the standard deviation of the angle of W1 and W5 we see the surprising result that W1 is *less* sensitive, for the two bootstrap-based robustness procedures, than W5. However,

FIG. A2. As in Fig. 2, but for synthetic data that aims to simulate convective rainfall over a large domain. Rainfall amounts at each location are simulated using uniform distributions with a range from 0 to 4. The two locations are considered to be far enough apart that no single rain event would cause rain at both locations simultaneously.

we have already discussed in section 2 why the bootstrap procedures may be flawed when considering the sensitivity of W1, which could explain these results. When using the MVN-uncertainty procedure, W1 is the least robust, followed by W5, followed by DCA, with all differences being significant. However, as the MVN-sensitivity procedure can also give misleading results when the ensemble data are not really normally distributed, a final statement on whether W1 or W5 is more robust in terms of angle cannot be made based solely on these results, since the three procedures disagree on the ranking.

Once again we have repeated the analysis in this section for an ensemble size reduced to 10 (Fig. C3). The results are qualitatively the same as the results for the 50-member ensemble i.e., the DCA patterns are overall the most robust. Notably, even though W5 and DCA5 look very similar in Figs. C1 and C2, DCA5 is clearly more robust in Fig. C3b.

SENSITIVITY TO THE DOMAIN

We next look at another aspect of the robustness of the results, which is the sensitivity of the worst-case scenario methods to the exact size of the domain. Figure 7 shows the distribution of the standard deviation of a and α over all forecast cases. In terms of amplitude, W1 and DCA1 have a higher uncertainty than W5 and DCA5. In terms of mean angle standard deviation, W1 and W5 are very similar, but W1 has a higher case-to-case variability. DCA shows a lower sensitivity than the other two methods (Fig. 7b). The dominance of DCA in this respect was unanticipated, but is an additional reason to favor it over the other two methods.

4. Discussion and conclusions

Our goal in this study is to compare possible approaches for providing information on spatial worst-case scenarios to ensemble forecast users who are unable to analyze every ensemble member individually. We specifically consider the idea of presenting an ensemble as the ensemble mean and a single additional pattern that illustrates possible worst-case deviations from the ensemble mean. The challenge is then how to generate the worst-case pattern. We have compared four methods to extract plausible worst-cases from a spatial ensemble weather forecast: taking the worst ensemble member (W1), taking the mean of the five worst members (W5), and two versions of directional component analysis (DCA). DCA shows a lower sensitivity than the other two methods (Fig. 7b). The dominance of DCA in this respect was unanticipated, but is an additional reason to favor it over the other two methods.

FIG. B1. Verifying analysis anomaly (ERA5 reanalysis with respect to forecast ensemble mean) for the forecasts shown in Figs. 3a and 4b.

FIG. C1. As in Fig. 3, but for an ensemble reduced to 10 randomly selected members.

the mean of the five worst members (DCA5). This allows us to focus on differences in the patterns. We have first tested the different methods on synthetic data, and then applied them to temperature ensemble forecast data over Europe from ECMWF. In all tests, the four methods generate different worst-case scenarios from the same forecast.

In the synthetic data tests, W1 is highly variable from one ensemble realization to the next, while DCA1, which has the same amplitude, shows much less variability. The theoretical grounding of DCA ensures that the DCA1 pattern also has a higher likelihood, if the statistical assumptions underlying its derivation are satisfied by the forecast data. W5 is less variable than W1, while DCA5 is somewhat less variable again, although the reduction in variability is now smaller.

In the real forecast data tests, we have first considered two cases in detail. In both cases, W1 shows large-amplitude, small-scale variability that does not appear in the patterns generated by the other methods, and that we doubt is robust. W5 is smoother than W1, and the corresponding DCA patterns are in turn smoother than both W1 and W5. We have also used a set of 18 real forecasts to test the three approaches for statistical

robustness, with three different procedures for estimating the robustness: bootstrapping, selecting a random subensemble, and fitting a multivariate normal (MVN) distribution to the ensemble and drawing a new ensemble from the MVN. DCA is overall the most robust approach. Whether the W1 or W5 pattern is more robust cannot be answered conclusively from this part of our study, as the different uncertainty estimation procedures give different rankings. This appears to be related to shortcomings in the procedures we have used for estimating robustness for the real example. We have also considered the robustness of the worst-case scenarios to small variations in the geographical domain. In these tests, DCA was more robust than the other methods.

a. Limits of applicability

There are various limitations of the methods we have discussed. All methods are limited by the fact that using a single worst-case cannot illustrate multimodality. We can then consider whether the patterns from the different methods are physically realistic. W1 should always be physically realistic—at least as realistic as the forecast model. However, the other methods may generate unrealistic patterns even if the

FIG. C2. As in Fig. 4, but for an ensemble reduced to 10 randomly selected members.

forecast model is realistic. This is because they involve averaging of some or all of the ensemble members. This can also be a problem with using the ensemble mean, especially in cases where the ensemble spread is large (WMO 2012).

For elliptically distributed data (such as multivariate normal or multivariate t distributed data) DCA possesses a number of optimality properties. However, it is still reasonable to apply DCA to ensembles which are not elliptically distributed. We have analyzed a synthetic example based on highly skewed gamma distributions in which the DCA methods are still more robust than W1 or W5. The factor that decides whether or not DCA and W5 should be used is not the shape of the marginal distribution, but relates to dependencies between locations, and whether or not averaging together ensemble members creates patterns which are plausible members themselves. One example in which averaging together ensemble members may not create patterns which are plausible members themselves is the case of convective rainfall. We can imagine a situation in which convective rainfall is forecast in different locations by different ensemble members, and we have provided a synthetic example to illustrate this case. If the members are averaged, localized strong rainfall will be replaced by distributed weak rainfall. For such a forecast, neither the ensemble mean, nor

W5, nor DCA, would be a realistic pattern. Even W1 would not be very useful: it would show heavy rainfall but in just one location. One can question the utility of using a single worst-case pattern at all in this case, and probability of precipitation products are likely more useful. Another example in which averaging together ensemble members might not create patterns which are plausible ensemble members themselves is for wind forecasts in certain situations, such as strong depressions, or tornadoes. Once again, in these situations the ensemble mean, W5 and DCA are not likely to be useful.

Another potential caveat of DCA is that when it is applied over large areas, there are likely to be spurious correlations of anomalies due to sampling variability, especially for grid points that are separated by long distances. It is possible that DCA could therefore be improved with covariance localization, which has been proposed for estimating covariance matrices of finite ensembles (Hamill et al. 2001).

b. Future developments

In this work, we have defined “extreme” with respect to the ensemble mean. All our methods could also be used with different definitions of extreme. One obvious candidate would be

FIG. C3. As in Fig. 6, but for an ensemble reduced to 10 randomly selected members.

extreme with respect to climatology. Another would be to use weights in space.

Finally, in this work, we have not attempted to validate our worst-case scenario forecasts. We have assumed that the ensembles have already been validated and have focused on testing methods for extracting statistical information from a validated ensemble. Prior to using these methods in practice, however, some form of validation would be desirable. Given that the worst-case scenarios have by definition low probability, a large number of past forecasts and the resources to run very large ensembles would be necessary for this. Another challenge would be that the validation would need to assess the probabilities of spatial patterns, which is a complex topic.

To conclude, we have shown that for certain kinds of ensemble DCA can provide plausible worst-case scenarios that are much more robust in various ways than the worst member of the ensemble and somewhat more robust than the mean of the five worst members. To the extent that the covariance matrix of the ensemble well represents the variability in the ensemble, DCA also gives a more likely but just as severe pattern relative to the other approaches. DCA, therefore, seems to be a good candidate for providing worst-case scenarios in the context of presenting an ensemble forecast as a mean and a worst-case, for those ensembles for which averaging ensemble members together makes sense.

Acknowledgments. SJ developed the initial ideas that led to this study and conducted the analysis with the synthetic data. SS conducted the analysis on the real forecast data and drafted the manuscript. All authors designed the study, discussed the results, and helped in improving the manuscript. SS was funded by the Department of Meteorology of Stockholm University. GM was partly supported by the Swedish Research Council Vetenskapsrådet (Grant 2016-03724).

Data availability statement. The code developed for will be openly available upon publication in the zenodo repository

with reserved doi:10.5281/zenodo.4282626 under <https://zenodo.org/record/4282626>. The ECMWF forecast data can be obtained from ECMWF's MARS archive.

APPENDIX A

Additional Synthetic Data

We show the results of analyses conducted on synthetic datasets that follow gamma and uniform distributions (Figs. A1 and A2). The figures are discussed in the main text.

APPENDIX B

Verifying Anomalies

We show the verifying anomalies from the ERA5 reanalysis for the example forecasts discussed in the main text (Fig. B1). The figure is discussed in the main text.

APPENDIX C

Results with Reduced Ensemble Size

The ECMWF ensemble forecast are comprised of 50 ensemble members. Many other operational ensemble forecast models, however, have smaller ensemble sizes. Here, we assess how well our methods are suited for smaller ensembles. For this, we reduce the ECMWF forecasts from

TABLE D1. The p values for difference of amplitude a between the worst member (W1) and the mean of the five worst members (W5), estimated with the one-sample t test.

Sensitivity procedure	$p(a)$
bootstrap	1.8×10^4
bootstrap_subens_25	5.0×10^5
mvn_uncertainty	7.0×10^{10}

TABLE D2. The p values for difference of angle α , estimated with the one-sample t test.

Method 1	Method 2	Sensitivity procedure	p (α)
worst_member (W1)	dca1	bootstrap	0.120
worst_member (W1)	worst_5 (W5)	bootstrap	0.053
dca5	worst_5 (W5)	bootstrap	0.000
worst_member (W1)	dca1	bootstrap_subens_25	0.372
worst_member (W1)	worst_5 (W5)	bootstrap_subens_25	0.442
dca5	worst_5 (W5)	bootstrap_subens_25	0.002
worst_member (W1)	dca1	mvn_uncertainty	0.000
worst_member (W1)	worst_5 (W5)	mvn_uncertainty	0.000
dca5	worst_5 (W5)	mvn_uncertainty	0.000

50 to 10 members, via only using the first 10 members. Since the members do not have any particular order, this results in a random subensemble. The bootstrapping and the subensemble uncertainty estimation procedure were adapted to the smaller ensemble size, so that the subensemble is made of 5 instead of 25 members. The W5 method is still defined as the mean of the worst five members, which is now half the ensemble.

The results are shown in Figs. C1–C3. All worst-case scenarios are less extreme than for the full ensemble (cf. Figs. 3, 4 and C1, C2). This is not surprising, and is a direct cause of the smaller sample size. In terms of robustness, the results are similar to the full ensemble. For the MVN uncertainty estimation procedure, DCA clearly is the most robust (small standard deviation of α). The two other uncertainty estimation procedures are probably unreliable, because the sample size is very small (10), and the caveats mentioned in section 2c(1). for the full ensemble are even more problematic here.

APPENDIX D

Results of Statistical Tests

The results of statistical tests comparing the robustness of a and α for the different worst-case scenario methods are shown in Tables D1 and D2.

REFERENCES

- Chu, P. C., S. E. Miller, and J. A. Hansen, 2015: Fuel-saving ship route using the Navy's ensemble meteorological and oceanic forecasts. *J. Def. Model. Simul.*, **12**, 41–56, <https://doi.org/10.1177/1548512913516552>.
- ECMWF, 2015: ENS cluster products. ECMWF, Reading, United Kingdom, accessed 28 June 2021, <https://www.ecmwf.int/en/forecasts/documentation-and-support/ens-cluster-products>.
- Ferranti, L., and S. Corti, 2011: New clustering products. *ECMWF Newsletter*, No. 127, Reading, ECMWF, United Kingdom, 6–11, <https://doi.org/10.21957/lr3bcise>.
- Fouillet, A., and Coauthors, 2006: Excess mortality related to the August 2003 heat wave in France. *Int. Arch. Occup. Environ. Health*, **80**, 16–24, <https://doi.org/10.1007/s00420-006-0089-4>.
- Fundel, V. J., N. Fleischhut, S. M. Herzog, M. Göber, and R. Hagedorn, 2019: Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users. *Quart. J. Roy. Meteor. Soc.*, **145**, 210–231, <https://doi.org/10.1002/qj.3482>.
- Hamill, T. M., J. S. Whitaker, and C. Snyder, 2001: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.*, **129**, 2776–2790, [https://doi.org/10.1175/1520-0493\(2001\)129<2776:DFOBE>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<2776:DFOBE>2.0.CO;2).
- Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, <https://doi.org/10.1002/qj.3803>.
- Hoffschmidt, M., J.-R. Bidlot, B. Hansen, and P. A. E. M. Janssen, 1999: Potential benefit of ensemble forecasts for ship routing. ECMWF Tech. Memo. 287, 25 pp., <https://doi.org/10.21957/ucgxanos0>.
- Jewson, S., 2020: An alternative to PCA for estimating dominant patterns of climate variability and extremes, with application to U.S. and China seasonal rainfall. *Atmosphere*, **11**, 354, <https://doi.org/10.3390/atmos11040354>.
- Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Roy. Meteor. Soc.*, **129**, 3037–3057, <https://doi.org/10.1256/qj.02.152>.
- Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539, <https://doi.org/10.1016/j.jcp.2007.02.014>.
- Molinder, J., H. Körnich, E. Olsson, H. Bergström, and A. Sjöblom, 2018: Probabilistic forecasting of wind power production losses in cold climates: A case study. *Wind Energy Sci.*, **3**, 667–680, <https://doi.org/10.5194/wes-3-667-2018>.
- Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliaigis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, <https://doi.org/10.1002/qj.49712252905>.
- Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.*, **128**, 747–774, <https://doi.org/10.1256/0035900021643593>.
- Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, <https://doi.org/10.1002/qj.49712656313>.
- Rossa, A., K. Liechti, M. Zappa, M. Bruen, U. Germann, G. Haase, C. Keil, and P. Krahe, 2011: The COST 731 Action: A review on uncertainty propagation in advanced hydro-meteorological forecast systems. *Atmos. Res.*, **100**, 150–167, <https://doi.org/10.1016/j.atmosres.2010.11.016>.

- Stensrud, D. J., and N. Yussouf, 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.*, **131**, 2510–2524, [https://doi.org/10.1175/1520-0493\(2003\)131<2510:SEPOMT>2.0.CO;2](https://doi.org/10.1175/1520-0493(2003)131<2510:SEPOMT>2.0.CO;2).
- Taylor, J. W., P. E. McSharry, and R. Buizza, 2009: Wind power density forecasting using ensemble predictions and time series models. *IEEE Trans. Energy Convers.*, **24**, 775–782, <https://doi.org/10.1109/TEC.2009.2025431>.
- Vié, B., G. Molinié, O. Nuissier, B. Vincendon, V. Ducrocq, F. Bouttier, and E. Richard, 2012: Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events. *Nat. Hazards Earth Syst. Sci.*, **12**, 2631–2645, <https://doi.org/10.5194/nhess-12-2631-2012>.
- Wasserman, L., 2004: *All of Statistics*. Springer, 442 pp.
- Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.
- WMO, 2012: Guidelines on ensemble prediction systems and forecasting. WMO-1091, 23 pp., http://www.wmo.int/pages/prog/www/Documents/1091_en.pdf.