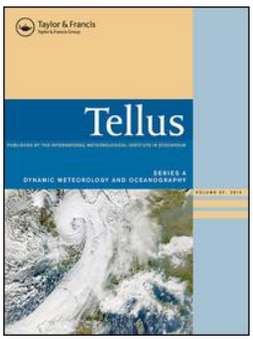

Tellus A: Dynamic Meteorology and Oceanography

ISSN: (Print) 1600-0870 (Online) Journal homepage: https://www.tandfonline.com/loi/zela20

How well can an ensemble predict the uncertainty in the location of winter storm precipitation?

Fan Han & Istvan Szunyogh

To cite this article: Fan Han & Istvan Szunyogh (2018) How well can an ensemble predict the uncertainty in the location of winter storm precipitation?, Tellus A: Dynamic Meteorology and Oceanography, 70:1, 1-10, DOI: 10.1080/16000870.2018.1440870

To link to this article: https://doi.org/10.1080/16000870.2018.1440870

Published online: 27 Feb 2018.

Tellus SERIES A DYNAMIC METEOROLOGY AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

How well can an ensemble predict the uncertainty in the location of winter storm precipitation?

By FAN HAN* and ISTVAN SZUNYOGH, *Department of Atmospheric Sciences, Texas A&M University, College Station, TX, USA*

(Manuscript received 3 October 2017; in final form 11 February 2018)

ABSTRACT

A pair of morphing-based ensemble forecast diagnostics is proposed for the verification of the location of precipitation events. The diagnostics are applied to operational global ensemble forecasts of winter storms in the United States in the winters of 2014/2015 and 2015/2016. A slowly developing systematic error is found to lead to an unrealistically fast eastward propagation of the storms in the week-two forecasts. Apart from this systematic error, the forecasts predict the uncertainty in the location of the precipitation events reliably. They, however, also grossly underestimate the uncertainty of the amount of precipitation in the short (shorter than 5 days) forecast range.

Keywords: ensemble verification, precipitation, morphing, winter storm

1. Introduction

The predictability of a chaotic dynamical system is measured by the temporal growth of the magnitude of the errors in predictions of the system. For an Eulerian scalar state variable of a spatio-temporally chaotic system, the standard measure of the magnitude of the prediction error is the root-mean-square (rms) error, with the mean taken over the spatial domain of the system. The use of the rms error as the measure of prediction error, however, is problematic for a scalar state variable of sharp gradients, because for such a variable, the rms error indicates a rapid loss of predictability once the dominant features of the field become slightly misplaced. Intuition suggests that a proper error measure should indicate that the error is a small displacement of the dominant features. More generally, the measure should provide information about the magnitude of the displacement error, and also the errors in the amplitude and spatial structure of the dominant features.

An example for a scalar field of the aforementioned type is the precipitation field associated with an extratropical or tropical cyclone, whose evolution is driven by the spatio-temporally chaotic, multi-scale dynamics of the atmosphere, which organizes it into bands with sharp boundaries and a rich and rapidly changing structure within the bands (Fig. 1). If the precipitation bands are slightly misplaced in a forecast, the rms error indicates poor forecast quality, even if the precipitation field is otherwise well predicted.

The error in the prediction of a precipitation event can be characterized, at minimum, by three error components: the errors in the location, amplitude (amount) and structure of the predicted precipitation (e.g. Wernli et al., 2008). Motivated by the work of Keil and Craig (2007, 2009) on morphing-based precipitation verification techniques and a series of papers on digital image quality measures (Wang and Bovik, 2002; Wang et al., 2004; Wang and Bovik, 2009), we have developed a technique to estimate the three error components in deterministic precipitation forecasts (Han and Szunyogh, 2016, 2017). The goal of the present study is to extend our technique for the estimation of the location error component to ensemble forecasts. In particular, we derive diagnostics for the estimation of the systematic location error and the verification of the "spread-skill relationship" (e.g. Buizza, 1997) for the location. We apply the two diagnostics to operational global ensemble forecasts of the 32 United States winter storms that were named by The Weather Channel in the winters of 2014/2015 and 2015/2016.[1] We note that a recent study (Greybush et al., 2017) based on the examination of forecasts of two of the storms from the same winters showed the importance of using the ensemble approach for the prediction of winter storms.

2. Methodology

We assume that the location of a precipitation event can be described by a two-dimensional vector of location $\boldsymbol{r}$. While the verification technique we propose does not require the knowledge or estimation of $\boldsymbol{r}$, the assumption that the position of the event can be described by a single location $\boldsymbol{r}$ makes its justification more transparent.[2] Because we consider $\boldsymbol{r}$ a random

*Corresponding author. e-mail: hanfan5598@gmail.com

Citation: Tellus A: 2018, **70**, 1440870, https://doi.org/10.1080/16000870.2018.1440870

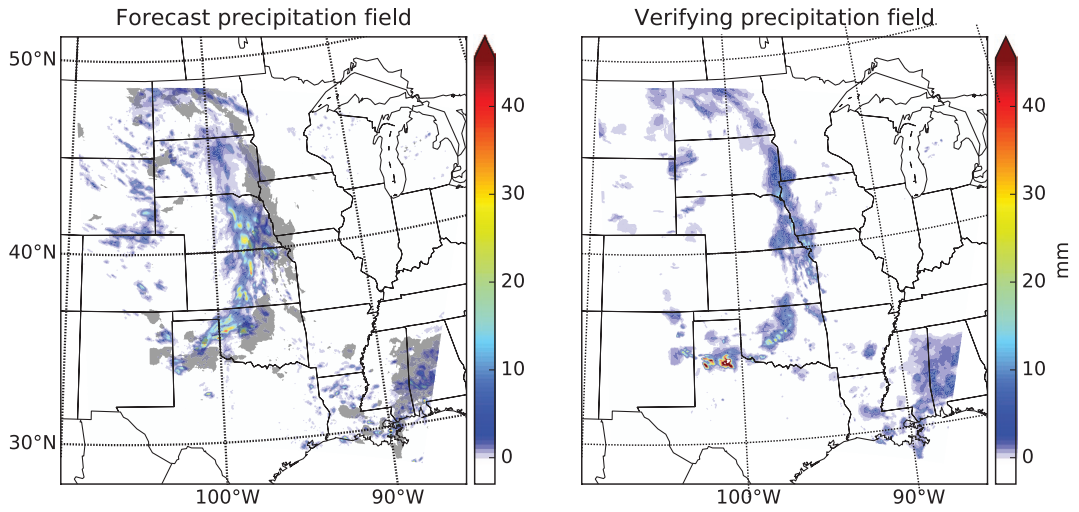

Fig. 1. An example of a slightly misplaced forecast precipitation field.
Notes: The forecast field (left) is the 24-h forecast of the 1-h accumulated precipitation starting at 0000 UTC 1 June 2005 and the verifying analysis field (right) is the 1-h accumulated Stage II precipitation analysis. In the left panel, the grey shades indicate the contours of the verifying precipitation field.

variable, the position $\boldsymbol{r}_a$ of the event in the verifying analysis is a realization of $\boldsymbol{r}$. Likewise, the positions $\boldsymbol{r}_f^k$ $(k = 1, 2 \ldots, K)$ of the event in the K forecast ensemble members are also realizations of $\boldsymbol{r}$.

Consider a set of verification cases, in which subsets of cases may be related to the same weather event at different verification times. Let M be the total number of verification cases. For each verification case $m(m = 1, 2, \ldots, M)$, we consider all ensemble forecasts from an archived data-set that are valid at the (verification) time of the case. While there are different lead time forecasts for each case, not all ensemble members predict a storm at each lead time. We therefore introduce the notation $K'\left(m, t_f\right), K'\left(m, t_f\right) \leq K$, for the number of ensemble members that at lead time t_f include a precipitation event that may be related to winter storm $m(m = 1, 2, \ldots, M)$. (We will give a formal definition of "may be related" shortly.) Our goal is to verify the ensemble-based prediction of the mean and standard deviation of the conditional probability distribution of $\boldsymbol{r}$ subject to the condition that the forecast verification feature may be related to an observed winter storm.

Our task has two parts. First, we have to identify the ensemble members that include a precipitation event that may be related to a verifying event. Second, we have to verify the ensemble-based estimates of the statistical parameters of the conditional probability distribution of $\boldsymbol{r}$ for the collection of cases that we identify. The only information available to us at the beginning of the process is the knowledge of the fields of verifying precipitation data $P^a(m)(m = 1, 2, \ldots, M)$ in a search region, which is selected such that the verifying event is at about its centre, and the related K-member ensembles of forecast precipitation fields $P^k\left(m, t_f\right)(k = 1, 2, \ldots, K)$.

We use the technique of Han and Szunyogh (2017) to first find a shift vector $d\boldsymbol{X}^k(m, t_f) = \left(dU^k(m, t_f), dV^k(m, t_f)\right)$ for each ensemble member k, forecast case m and lead time t_f that corrects the location error. ($dU^k(m, t_f)$ and $dV^k(m, t_f)$ are the zonal and meridional component of $d\boldsymbol{X}^k(m, t_f)$, respectively.) Formally, the shift vector that "corrects the location error" is the vector that shifts $P^k(m, t_f)$ such that it maximizes the similarity between P^a(m) and the shifted $P^k(m, t_f)$ field, $P^k_{shift}(m, t_f)$. We think of $d\boldsymbol{X}^k(m, t_f)$ as the difference between the location $\boldsymbol{r}_a(m)$ of the verifying precipitation feature and the predicted location $\boldsymbol{r}_f^k(m, t_f)(k = 1, 2, \ldots, K)$ of the same feature in ensemble member k, that is,

$$d\boldsymbol{X}^k\left(m, t_f\right) = \boldsymbol{r}_a(m) - \boldsymbol{r}_f^k\left(m, t_f\right),\ k = 1, 2, \ldots, K. \tag{1}$$

We measure the similarity between $P^a(m)$ and $P^k_{shifted}(m, t_f)$ by the *Amplitude and Structural Similarity Index Measure (ASSIM)* (Han and Szunyogh, 2017), which is an adaptation of the Structural Similarity Index Measure (SSIM) of Wang et al. (2004) and Wang and Bovik (2009). We choose the free parameters of the measure such that it gives equal weights to the similarity of the amplitude, the similarity of the spatial variability and the point-wise correlation of the two fields. ASSIM takes a value in the closed interval [0, 1], with one indicating identical fields and a lower value indicating less similarity between the two fields. We assume that the precipitation feature of ensemble member k "may be related to" winter storm m, if ASSIM for P^a(m) and $P^k_{shift}(m, t_f)$ is equal to, or larger than a prescribed threshold value a. $K'\left(m, t_f\right)$ therefore is the number of ensemble members for forecast case $m(m = 1, 2, \ldots, M)$ and forecast lead time t_f for which ASSIM is larger than a.

We expect $K'(m,t_f)$ to be a monotonically decreasing function of the forecast lead time t_f and require that $K'(m,t_f) \geq 2$ for all forecast cases used in the computation of the diagnostics at t_f. We denote the number of forecast cases that satisfy the latter condition by $M'(t_f)$. Estimates of the statistics based on such small ensemble sizes can be included in the diagnostics, because while the sampling errors (the estimation errors due to a small ensemble) can be large for a particular case m and lead time t_f, the expected value and the standard deviation of the sampling errors are known from the theory of statistics even for such small sample sizes. In particular, if the ensemble members $\boldsymbol{r}_f^k (k = 1, 2, \ldots, K')$ sample the *true distribution* of $\boldsymbol{r}$, the ensemble average

$$\overline{\boldsymbol{r}_f} = \frac{1}{K'} \sum_{k=1}^{K'} \boldsymbol{r}_f^k \tag{2}$$

estimates the (unknown) true mean $\bar{\boldsymbol{r}} = E[\boldsymbol{r}]$ of the distribution with an error

$$\boldsymbol{b}_{loc} = \bar{\boldsymbol{r}} - \overline{\boldsymbol{r}_f}, \tag{3}$$

whose mean is

$$E[\boldsymbol{b}_{loc}] = 0, \tag{4}$$

and mean-square is

$$E\left[\left(\boldsymbol{b}_{loc}\right)^2\right] = E\left[\left(\boldsymbol{b}_{loc} - E\left[\boldsymbol{b}_{loc}\right]\right)^2\right] = \frac{1}{K'}\Sigma_{loc}^2. \tag{5}$$

Here,

$$\Sigma_{loc}^2 = E\left[(\boldsymbol{r} - \bar{\boldsymbol{r}})^2\right] \tag{6}$$

is the (unknown) true variance of $\boldsymbol{r}$. Hereafter, $E[\cdot]$ denotes the expected value of the random variable in the brackets. In our proposed diagnostics, this expected value is estimated by an average over the $M'(t_f)$ verification cases.

Taking the ensemble mean of Equation (1) yields

$$\overline{d\boldsymbol{X}}(m,t_f) = \boldsymbol{r}_a(m) - \overline{\boldsymbol{r}_f}(m,t_f), \tag{7}$$

where

$$\overline{d\boldsymbol{X}}(m,t_f) = \frac{1}{K'(m,t_f)} \sum_{k=1}^{K'} d\boldsymbol{X}^k(m,t_f). \tag{8}$$

According to Equation (7), $\overline{d\boldsymbol{X}} = \left(\overline{dU}, \overline{dV}\right)$ is the difference between a realization $\boldsymbol{r}_a$ of $\boldsymbol{r}$ and the prediction $\overline{\boldsymbol{r}_f}$ of the mean $\bar{\boldsymbol{r}}$. Equation (7) can also be written in the equivalent form

$$\overline{d\boldsymbol{X}}(m,t_f) = \boldsymbol{\epsilon}_{loc}(m,t_f) + \boldsymbol{b}_{loc}(m,t_f), \tag{9}$$

where

$$\boldsymbol{\epsilon}_{loc}(m,t_f) = \boldsymbol{r}_a(m) - \bar{\boldsymbol{r}}(m,t_f) \tag{10}$$

is a realization of the random variable $\boldsymbol{r} - \bar{\boldsymbol{r}}$, which we call the *location uncertainty*. Notice that Σ_{loc}^2 describes the magnitude of the location uncertainty (see Equation 6).

Ideally, the ensemble should sample the true probability distribution of the forecast variables given all sources of forecast uncertainty. Because this property cannot be verified directly (Talagrand et al., 1999), ensemble verification techniques examine necessary conditions for it. We follow this approach by deriving diagnostic equations that the ensemble forecasts would satisfy at forecast lead time t_f, if the ensemble sampled the true probability distribution of the forecast uncertainty.

2.1. *Location bias*

Because $\boldsymbol{r}_a$ is a realization of $\boldsymbol{r}$,

$$E\left[\boldsymbol{\epsilon}_{loc}\right](m,t_f) = E\left[\boldsymbol{r}_a - \bar{\boldsymbol{r}}\right](m,t_f) = \boldsymbol{0}, \tag{11}$$

and the expected value of Equation (9) is

$$E\left[\boldsymbol{b}_{loc}\right](m,t_f) = E\left[\overline{d\boldsymbol{X}}\right](m,t_f). \tag{12}$$

Hence, the estimate

$$\boldsymbol{\beta}_{loc}(t_f) = \frac{1}{M'(t_f)} \sum_{\substack{m=1 \\ K'(m,t_f) \geq 2}}^{M} \overline{d\boldsymbol{X}}(m,t_f) \tag{13}$$

of the right-hand side of Equation (12) is also an estimate of the location bias $E\left[\boldsymbol{b}_{loc}\right](t_f)$. $\boldsymbol{\beta}_{loc}(t_f)$ is an unbiased estimate of the location bias, because Equation (4) also applies if $\boldsymbol{b}_{loc}$ is replaced by $\overline{d\boldsymbol{X}}(m,t_f)$.

2.2. *Spread–skill relationship*

The spread–skill relationship diagnostic of ensemble forecasting takes advantage of the statistical relationship between the ensemble variance and the difference between the verifying data and the ensemble mean: the expected value of the ensemble variance and the expected value of the square of the difference between the verifying data and the ensemble mean are both estimates of Σ_{loc}^2. Our goal is to formally express this relationship with the help of the ensemble of shift vectors.

Making use of Equations (1) and (7) yields

$$\begin{aligned}\left(\boldsymbol{r}_f^k - \overline{\boldsymbol{r}_f}\right)^2(m,t_f) &= \left(\boldsymbol{r}_a - d\boldsymbol{X}^k + \overline{d\boldsymbol{X}} - \boldsymbol{r}_a\right)^2(m,t_f) \\ &= \left(d\boldsymbol{X}^k - \overline{d\boldsymbol{X}}\right)^2(m,t_f).\end{aligned} \tag{14}$$

Thus, the expected value of the ensemble variance of the location can be expressed by the expected value of the ensemble variance of the shift vectors as

$$E\left[\frac{1}{K'-1}\sum_{k=1}^{K'}\left(\boldsymbol{r}_f^k - \overline{\boldsymbol{r}_f}\right)^2\right](t_f) = E\left[\frac{1}{K'-1}\sum_{k=1}^{K'}\left(\overline{d\boldsymbol{X}} - d\boldsymbol{X}^k\right)^2\right](t_f). \tag{15}$$

The expected value of the square of the difference between the verifying data and the ensemble mean can be expressed with the help of the shift vectors by taking the expected value of the square of Equation (7), which leads to

$$E\left[\left(\boldsymbol{r}_a-\overline{\boldsymbol{r}_f}\right)^2\right]\left(t_f\right)=E\left[\overline{d\boldsymbol{X}}^2\right]\left(t_f\right). \tag{16}$$

In addition,

$$E\left[\left(\boldsymbol{r}_a-\overline{\boldsymbol{r}_f}\right)^2\right]\left(t_f\right)=E\left[\left(\boldsymbol{r}-\overline{\boldsymbol{r}_f}\right)^2\right]\left(t_f\right)=E\left[\left((\boldsymbol{r}-\bar{\boldsymbol{r}})+\boldsymbol{b}_{loc}\right)^2\right]\left(t_f\right). \tag{17}$$

Because at this point we assume that the estimation error $\boldsymbol{b}_{loc}$ of the mean $\bar{\boldsymbol{r}}$ is purely due to sampling errors, making use of Equations (5) and (6) leads to

$$E\left[\left((\boldsymbol{r}-\bar{\boldsymbol{r}})+\boldsymbol{b}_{loc}\right)^2\right]\left(t_f\right)=\Sigma_{loc}^2\left(t_f\right)+E\left[\left(\boldsymbol{b}_{loc}\right)^2\right]\left(t_f\right)=\frac{K'+1}{K'}\Sigma_{loc}^2\left(t_f\right). \tag{18}$$

First, combining Equations (17) and (18) and then making use of Equation (16) yields

$$\Sigma_{loc}^2\left(t_f\right)=E\left[\frac{K'}{K'+1}\left(\boldsymbol{r}_a-\overline{\boldsymbol{r}_f}\right)^2\right]\left(t_f\right)=E\left[\frac{K'}{K'+1}\overline{d\boldsymbol{X}}^2\right]\left(t_f\right). \tag{19}$$

Because the left-hand side of Equation (15) is a prediction-based estimate of Σ_{loc}^2, for a perfectly formulated ensemble, the right-hand side of Equation (15) would be equal to the right-hand side of Equation (19); that is, we would find that

$$E\left[\frac{1}{K'-1}\sum_{k=1}^{K'}\left(d\boldsymbol{X}^k-\overline{d\boldsymbol{X}}\right)^2\right]\left(t_f\right)=E\left[\frac{K'}{K'+1}\overline{d\boldsymbol{X}}^2\right]\left(t_f\right). \tag{20}$$

The left-hand side is the "ensemble spread" of the forecast location and the right-hand side is the deterministic "forecast skill" (mean-square error) of the ensemble mean forecast. The expected values in Equation (20) can be estimated by taking averages over the sample of M forecast cases. That is,

$$\sigma_{loc}^2\left(t_f\right)=\frac{1}{M'\left(t_f\right)}\sum_{\substack{m=1\\K'\left(m,t_f\right)\geq 2}}^{M}\left[\frac{1}{K'\left(m,t_f\right)-1}\sum_{k=1}^{K'\left(m,t_f\right)}\left(d\boldsymbol{X}^k\left(m,t_f\right)-\overline{d\boldsymbol{X}}\left(m,t_f\right)\right)^2\right] \tag{21}$$

is an estimate of the left-hand side and

$$\delta_{loc}^2\left(t_f\right)=\frac{1}{M'\left(t_f\right)}\sum_{\substack{m=1\\K'\left(m,t_f\right)\geq 2}}^{M}\left[\frac{K'\left(m,t_f\right)}{K'\left(m,t_f\right)+1}\overline{d\boldsymbol{X}}^2\left(m,t_f\right)\right] \tag{22}$$

is an estimate of the right-hand side of Equation (20). The factor $K'/(K'+1)$ on the right-hand side of Equation (20) may seem unusual, because in the ensemble forecasting literature the effects of sampling errors caused by the limited number of ensemble members on the spread–skill relationship is rarely considered. The omission of the normalization factor introduces only small errors into the estimate of the "skill" for the typical number (>20) of ensemble members (Fig. 2). What makes our situation special is that we allow for such small values of K' as 2, for which the normalization factor is $K'/(K'+1)=2/3\approx 0.67$, which is significantly smaller than 1.

There is one additional issue that can affect the spread–skill relationship in the specific case of our morphing based verification technique, even if the ensemble correctly samples the true probability distribution: because the size of the search region limits the magnitude of the location error that the technique of Han and Szunyogh (2017) can detect, $\boldsymbol{r}_f^k-\overline{\boldsymbol{r}_f}$ and $\boldsymbol{r}_a-\overline{\boldsymbol{r}_f}$ is not always able to fully sample the tails of the probability distribution of the forecast

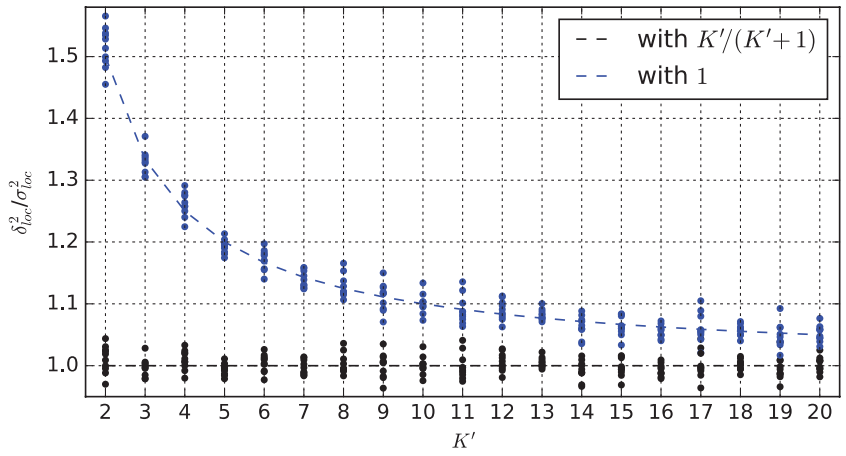

Fig. 2. Illustration of the effect of the normalization factor $K'/(K'+1)$ on the ratio of the estimates of the right- and left-hand sides of Equation (20). *Notes: The* values shown are based on randomly generated samples of 10, 000 realizations of $\boldsymbol{r}$, $\boldsymbol{r}_a$, $\boldsymbol{r}_f^k, k=1,\ldots,K'$, assuming a Gaussian random distribution with mean $\bar{\boldsymbol{r}}=\boldsymbol{0}$. Because of the sampling fluctuations, the simulation is repeated 10 times for each value of K' and the results are shown by scatter-plots.

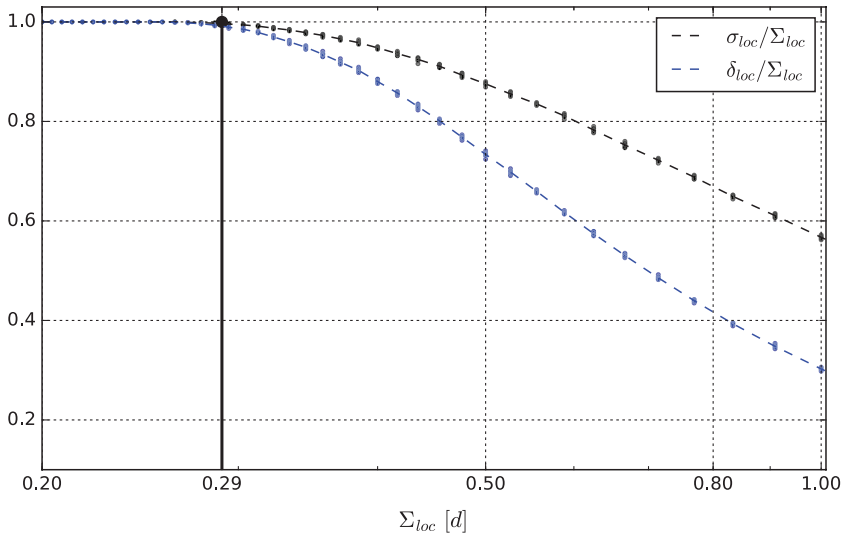

Fig. 3. The dependence of the prediction σ_{loc} and the estimate δ_{loc} on Σ_{loc} for a finite size search region.
Notes: The values shown are based on a randomly generated sample of 1000 realizations of $\boldsymbol{r}$, $\boldsymbol{r}_a$, $\boldsymbol{r}_f^k, k = 1, \ldots, 100$, assuming a truncated Gaussian random distribution with mean $\bar{\boldsymbol{r}} = \boldsymbol{0}$ in a one-dimensional search domain of size $2d$. The values of Σ_{loc} on the x-axis are shown using the search radius d as unit, while the values of σ_{loc} and δ_{loc} on the y-axis are normalized by Σ_{loc}.

uncertainty $\boldsymbol{r} - \bar{\boldsymbol{r}}$. This factor becomes important for the longer forecast times, at which Σ_{loc}^2 is typically large. The potential effects of this problem can be investigated by generating random samples of $\boldsymbol{r}$, $\boldsymbol{r}_a$ and $\boldsymbol{r}_f^k$ for a prescribed value of Σ_{loc}^2 and then verifying the sample based estimates of Σ_{loc}^2 by δ_{loc}^2 and σ_{loc}^2. Figure 3 summarizes the results of a simulation experiment along this line for an idealized, one-dimensional search domain (the position vectors are scalars along a line) of length $2d$, assuming that $\boldsymbol{r}_a$ is at the centre of the domain. The probability distribution of the location is assumed to be Gaussian, except that the tails of the distribution are truncated at plus and minus two standard deviations of the Gaussian distribution. The figure shows that once Σ_{loc} is larger than a critical value (about $0.29d$), both σ_{loc} and δ_{loc} start to underestimate Σ_{loc}, but the magnitude of the estimation error grows faster for δ_{loc} than σ_{loc} as Σ_{loc} increases. We will make use of this finding in the analysis of the results on winter storms.

We have hitherto discussed the spread–skill relationship under the assumption that the ensemble sampled the true probability distribution of the forecast uncertainty of the location. To relax this assumption, it is important to first notice that the spread–skill relationship depends on the quality of the prediction of not only the variance, but also the mean of the probability distribution. In fact, as we have already discussed, sampling errors in the estimate of the mean have an effect on the spread–skill relationship. The other form of error in the prediction of the mean that we can account for in the spread–skill relationship is the forecast bias. In the presence of bias $\boldsymbol{\beta}_{loc} \neq \boldsymbol{0}$, Equation (22) can be replaced by

$$\delta_{loc}^2(t_f) = \frac{1}{M'(t_f)} \sum_{\substack{m=1 \\ K'(m,t_f) \geq 2}}^{M} \left[\frac{K'(m,t_f)}{K'(m,t_f)+1} \overline{d\boldsymbol{X}}^2(m,t_f) \right] - \left(\frac{1}{M'(t_f)} \sum_{\substack{m=1 \\ K'(m,t_f) \geq 2}}^{M} \left[\sqrt{\frac{K'(m,t_f)}{K'(m,t_f)+1}} \overline{d\boldsymbol{X}}(m,t_f) \right] \right)^2. \tag{23}$$

This equation can be obtained by noticing that in the presence of bias, Equation (5) becomes

$$E\left[\left(\boldsymbol{b}_{loc} \right)^2 \right] = E\left[\left(\boldsymbol{b}_{loc} - E\left[\boldsymbol{b}_{loc} \right] \right)^2 \right] = \frac{1}{K'} \Sigma_{loc}^2 + E^2\left[\boldsymbol{b}_{loc} \right], \tag{24}$$

thus leading to the modified version

$$E\left[\frac{1}{K'-1} \sum_{k=1}^{K'} \left(\overline{d\boldsymbol{X}} - d\boldsymbol{X}^k \right)^2 \right](t_f) = E\left[\frac{K'}{K'+1} \overline{d\boldsymbol{X}}^2 \right](t_f) - E^2\left[\sqrt{\frac{K'}{K'+1}} \overline{d\boldsymbol{X}} \right](t_f) \tag{25}$$

of Equation (20). It is important to notice that Equation (23) accounts only for the systematic part of the error in the prediction

of the mean. Hence, flow dependent errors in the prediction of the mean can still lead to a breakdown of the modified spread–skill relationship (Equation 25).

2.3. Amplitude uncertainty

While the focus of the present paper is on the verification of the location forecasts, we also show verification results for the amplitude forecasts to contrast the behaviour of the diagnostics for the two forecast parameters. We describe the precipitation amount (amplitude) associated with a weather event by the areal mean μ of the precipitation in the verification domain, that is, by the ratio of the total precipitation in the verification domain and the area of the verification domain. Similar to the position $\boldsymbol{r}$, we treat μ as a random variable. Let $\mu_f^k(m,t_f)$ $\left(k=1,2,\ldots,K'(m,t_f)\right)$ be the areal mean of the precipitation in ensemble member k for forecast case m at forecast time t_f, and $\mu_a(m)$ the areal mean of the precipitation in the verifying analysis.

The bias $E\left[b_{amp}\right]$ of the areal mean precipitation (amplitude bias), which is the expected value of the difference

$$b_{amp}(m,t_f)=\bar{\mu}(m,t_f)-\overline{\mu_f}(m,t_f) \tag{26}$$

between the mean $\bar{\mu}=E[\mu]$ and its ensemble-based prediction $\overline{\mu_f}$, can be estimated by

$$\beta_{amp}(t_f)=\frac{1}{M'(t_f)}\sum_{\substack{m=1\\K'(m,t_f)\geq 2}}^{M} b_{amp}(m,t_f)\ . \tag{27}$$

In addition, by analogy to the arguments made earlier for the location uncertainty, the ensemble-based estimate of the amplitude uncertainty,

$$\sigma_{amp}^2(t_f)=\frac{1}{M'(t_f)}\sum_{\substack{m=1\\K'(m,t_f)\geq 2}}^{M}\left[\frac{1}{K'(m,t_f)-1}\sum_{k=1}^{K'(m,t_f)}\left(\mu_f^k(m,t_f)-\overline{\mu_f}(m,t_f)\right)^2\right], \tag{28}$$

and the analysed uncertainty,

$$\delta_{amp}^2(t_f)=\frac{1}{M'(t_f)}\sum_{\substack{m=1\\K'(m,t_f)\geq 2}}^{M}\left[\frac{K'(m,t_f)}{K'(m,t_f)+1}\left(\mu_a(m)-\overline{\mu_f}(m,t_f)\right)^2\right], \tag{29}$$

should be equal at any forecast time t_f.

3. Data

The diagnostics are applied to 0.5° × 0.5° resolution, global, 15-day long, twice daily, 20-member ensemble forecasts from the National Centers for Environmental Prediction (NCEP) of the US National Weather Service (NWS). The precipitation field at time t (e.g. 1200 UTC 2 January 2016) is defined by the accumulated precipitation for the 6-h period starting at time t. Because each of the 32 storms is present in multiple forecasts, the total number of verification cases considered is 133.

The search region for the estimation of the location error is 112 × 80 = 8960 grid points, which is about a region of 4900 km × 4400 km. The location of the search region is adjusted for each case such that the verifying precipitation feature is in the middle of the search region. Selecting a larger search region would eliminate the potential problem illustrated by Fig. 3, but the presence of multiple precipitation features in a larger search domain would also greatly complicate the implementation of the technique of Han and Szunyogh (2017).

The forecasts are verified against Stage IV precipitation analyses, which are based on radar and gauge observations over the US (Lin and Mitchell, 2005). These analyses are estimates of the rainfall accumulation for approximately 4 km × 4 km pixels. Since the precipitation system of winter storms often extends over the ocean, where no Stage IV data are available, we use 0.5° × 0.5° resolution, calibrated short-term (6-h) forecasts from the European Centre for Medium Range Forecasts (ECMWF) to fill the gaps in the verification data. Only those cases are included in the statistics for which more than 30% of the total precipitation in the verification region is associated with Stage IV data. The latter criterion reduces the number of forecast cases M from 133 to 83.

4. Results

4.1. The dependence of K' and M' on the forecast lead time

We start the examination of the results for the winter storms by an investigation of the typical number of ensemble members that predict the verifying storm. To be precise, we investigate the average of $K'(m,t_f), m=1,2,\ldots,M$, over the $M=83$ forecast cases (Fig. 4). As expected, the average K' rapidly decreases with forecast time t_f. In addition, the decrease is more rapid for the larger values of a, that is, when a higher degree of similarity is required between the forecast and the verifying precipitation to declare that they are likely to be related. The saturation value of the curves in the figure also strongly depends on a. Because the saturation of the curves indicates that K' no longer depends on the initial conditions, the saturation value of the ratio K'/K is an estimate of the likelihood that a storm is found sufficiently similar to the verifying storm by pure chance rather than due to forecast skill. This likelihood decreases with the increase of the

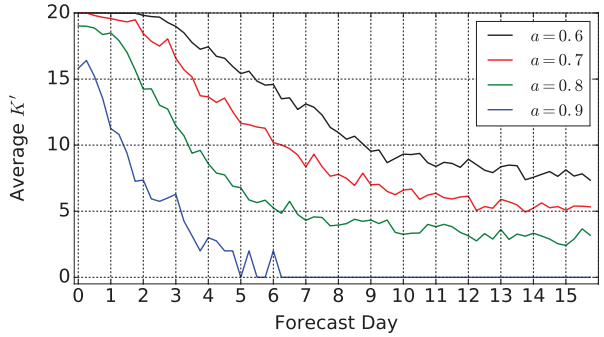

Fig. 4. Evolution of the average of K' over all forecast cases for different values of a in the range from 0.6 to 0.9.

required degree of similarity a between the forecast and the verifying precipitation event. In fact, the saturation value of K'/K can be reduced to zero by making the choice $a = 0.9$, but in that case, $K' < K$ even at initial time, which suggests that requiring such high degree of similarity between the forecast and the verifying precipitation field is unrealistic at the current level of modelling capabilities.

The decrease of the average of $K'(m, t_f)$ with forecast time t_f suggests that $M'(m, t_f)$ is also likely to decrease with the increase of t_f. This expectation is confirmed by the actual numbers (Fig. 5) that suggest that in order to maintain a sufficient sample size, a should be chosen not to be larger than $a = 0.7$. Because $a = 0.7$ is also the largest value of a for which $K' = K$ at analysis time (Fig. 4), we show the verification statistics for that particular value of a. We note, however, that we also carried out calculations for different values of a in the range from $a = 0.6$ to $a = 0.9$, and found that the verification results were robust to the choice of a, except for the presence of a higher level of noise for the larger values of a. We also note that Han and Szunyogh (2017) found the verification statistics to be similarly robust to the choice of a for deterministic forecasts.

4.2. *Location bias and uncertainty*

For the operational forecasts, both components of the estimated location bias are negligible for the first four forecast days (Fig. 6).

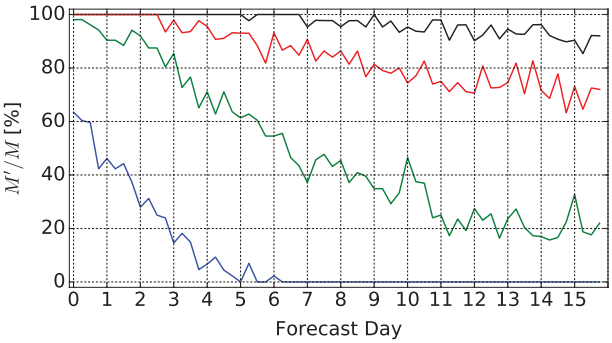

Fig. 5. Evolution of the percentage M'/M of the forecast cases for which $K' \geq 2$ for different values of a in the range from 0.6 to 0.9.

After that, the magnitude of the zonal component gradually increases until saturation at about forecast lead time day 9–10 at a level of about – 300 km. The negative sign indicates that the eastward propagation of the storms is faster in the forecasts than reality. This result is consistent with the findings of Herrera et al. (2016) and Loeser et al. (2017) that the operational ensemble forecasts of the leading prediction centres of the world have a slowly developing bias that results in an overly zonal large scale flow (a flow that does not have a realistic south–north meandering large-scale component) at the longer lead times in the forecasts: because the large-scale flow acts as a guide for the eastward propagating storms, the overly zonal large-scale flow leads to an unrealistically fast eastward propagation of the storms and their precipitation systems. This slowly developing forecast bias is an indication of systematic model errors.

Figure 7 shows the functions $\sigma_{loc}(t_f)$, $\delta_{loc}(t_f)$, and $\delta_{loc}(t_f)$ after bias correction. There is a good agreement between $\sigma_{loc}(t_f)$ and the bias-corrected $\delta_{loc}(t_f)$ up to about 6 days. Beyond that, σ_{loc} becomes larger than both δ_{loc} and the bias corrected δ_{loc}, which indicates that the sampling problem illustrated in Fig. 3 does indeed affect the results. The general shape of the curves indicates a rapid chaotic growth of the forecast uncertainty before reaching saturation at about t_f = 11–14 days. This is the (absolute) predictability time limit for the location of the storms, the time beyond which no storm can be predicted with an accuracy higher than the accuracy of a forecast based on climatology. It should be noted that the predictability time limit for a specific storm can be significantly shorter than the absolute predictability time limit (e.g. Greybush et al., 2017).

The saturation level of Σ, which is an estimate of the climatological value of the standard deviation of the distance between the locations of winter storms, can be estimated based on the results of Figs. 3 and 7. First, estimates of the saturation value of σ_{loc}, 1015 km, and the bias-corrected value of $\delta_{loc}(t_f)$, 626 km, can be obtained by a Lorenz-curve analysis (e.g. Han and Szunyogh, 2017). These estimates correspond to a ratio of 1015/626 = 1.62 of the values along the two curves in Fig. 3, which yields an estimate of Σ = 1530 km.

The information that the ratio between the saturation values of the diagnostics is 1.62 can also be used to obtain numerical estimates of the effective search radius d and the critical value $0.29d$: the x-value that corresponds to the ratio of 1.62 in Fig. 3 is $\Sigma = 0.81d$, which combined with the estimate 1530 km of Σ leads to d = 1886 km and $0.29d$ = 550 km. Notice that the estimate of d is somewhat smaller than half of the length of the verification region in either direction. This relation between the search radius and the size of the verification region is due to the property of the verification region that it has to include the entire forecast precipitation feature for an accurate estimation of the location error.

The good agreement between the three curves in Fig. 7 for the first six forecast days suggests that the ensemble forecasts provide accurate quantitative prediction of the uncertainty of

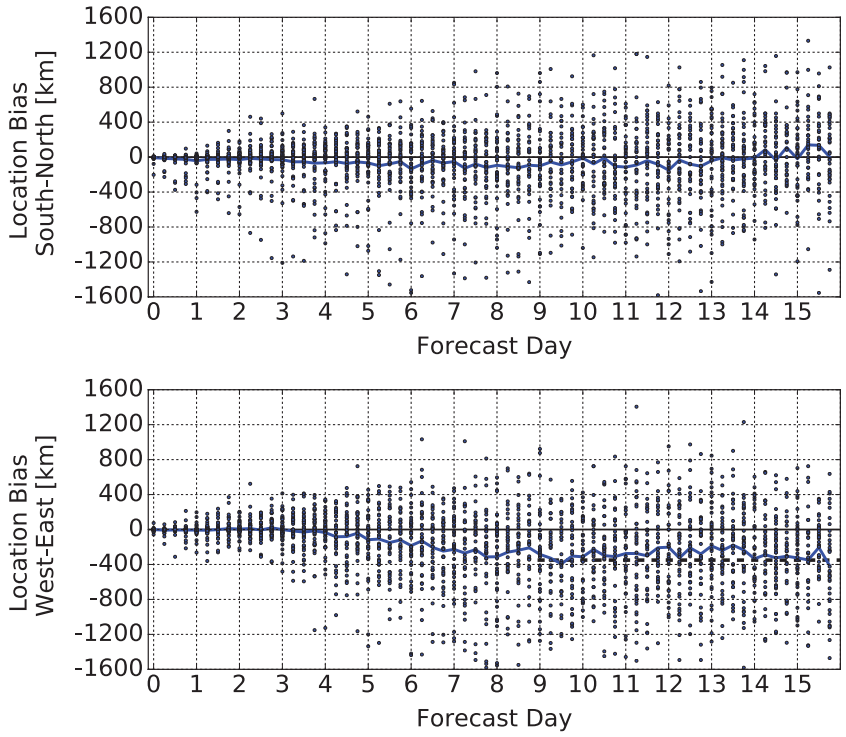

Fig. 6. Top: the values of $\overline{dV}$ for the individual ensemble forecasts (dark blue dots) and the evolution of the estimate of the meridional component of the location bias $E[\boldsymbol{b}_{loc}]$ with the forecast lead time, bottom: the values of $\overline{dU}$ for the individual ensemble forecasts (dark blue dots) and the evolution of the estimate of the zonal component of the location bias $E[\boldsymbol{b}_{loc}]$.

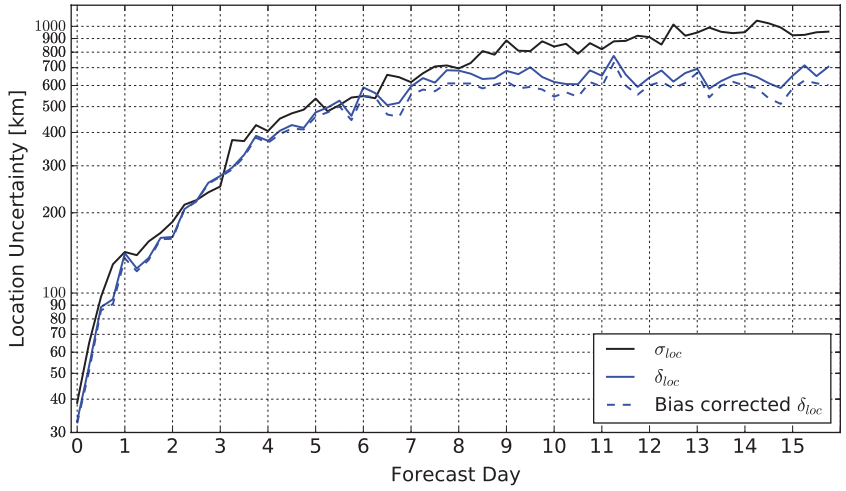

Fig. 7. The evolution of σ_{loc} (solid black), δ_{loc} (solid blue) and bias-corrected δ_{loc} (dashed blue) with the forecast lead time.

the location of the storms. Because the uncertainty in the location of the storms reflects uncertainty in the position of synoptic scale (~1000 km) features of the atmospheric flow, our results indicate that the ensemble is highly efficient in capturing the uncertainty at the synoptic scales. As a counter-example to this behaviour, next we show that the ensemble is much less successful in capturing the uncertainty in the multi-scale processes that determine the precipitation amount. These processes take place in the range from micrometres in the clouds to thousands of kilometres at the synoptic scales.

4.3. *Amplitude bias and uncertainty*

The evolution of the estimate of the amplitude bias β_{amp} (Fig. 8) indicates an initially rapidly and then slowly decreasing wet bias (over-prediction of the precipitation amount) in the

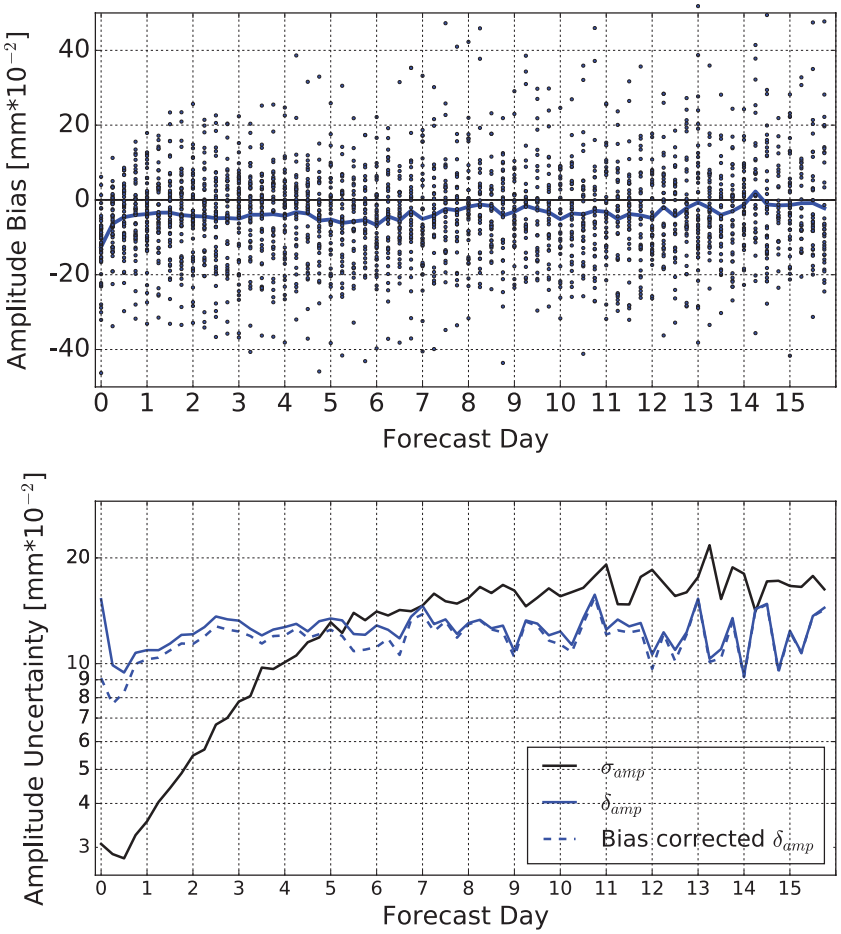

Fig. 8. Top: the values of $\mu_a - \overline{\mu_f}$ for the individual ensemble forecasts (dark blue dots) and the evolution of the estimate of the amplitude bias $E\left[b_{amp}\right]$ with the forecast lead time, bottom: the evolution of σ_{amp} (solid black), δ_{amp} (solid blue) and bias-corrected δ_{amp} (dashed blue) with the forecast lead time.

forecasts, which completely disappears after about 13 days. The strong initial adjustment in the precipitation, which indicates that the model initial conditions are in the basin of attraction of the "attractor of the model dynamics", but not exactly on the "attractor" (called "spin-up" in the numerical weather prediction jargon), has been a longstanding issue of numerical weather prediction. The fact that the wet bias eventually disappears as forecast time increases suggests that it is more likely to be due to problems with the initial conditions of the moist variables than systematic model errors. The relatively large initial value of $\delta_{amp}\left(t_f\right)$ also points to the initial conditions as the primary source of the amplitude forecast error in the first few forecast days. The ensemble spread $\sigma_{amp}\left(t_f\right)$, which predicts initially small and then rapidly growing amplitude errors, fails to capture the large initial errors, leading to an initially poor, but gradually improving ensemble performance. The fact that at the longer forecast lead times there is a much better general agreement between $\sigma_{amp}\left(t_f\right)$ and $\delta_{amp}\left(t_f\right)$ also supports the conclusion that the root of the initially large forecast errors and poor ensemble performance is primarily not a problem with the model climatology. This result calls for further research into the analysis and the generation of initial condition perturbations of moist model variables, for instance, by testing stochastic parameterization schemes and convective-allowing models in ensemble forecasting (e.g. Palmer et al., 2009; Khouider et al., 2010; Bengtsson and Körnich, 2016; Greybush et al., 2017).

5. Conclusions

In this paper, we introduced a morphing-based ensemble forecast verification technique for the location of precipitation events. We demonstrated the skill and the limitations of the technique with an application to operational ensemble forecasts of US winter storms. The results of this application suggest that the operational ensemble forecasts provide reliable forecasts of the uncertainty in

the location of the storms, except for a slowly developing systematic error that leads to an unrealistically fast eastward propagation of the storms in the week-two forecasts. We contrasted the good performance of the ensemble in predicting the uncertainty in storm location to its poor performance in predicting the uncertainty of the precipitation amount in the short (less than 5 days) forecast range.

The most important limitation of the proposed verification technique in its present form is that it treats all precipitation in the verification region as part of single precipitation system. This makes the careful selection of the verification region a critical part of the implementation of the technique and it may lead to spurious results in situations where there are multiple, equally important, isolated features in the verification region (e.g. precipitation associated with multiple isolated convective systems). In light of these limitations, we view the current version of the technique as a highly useful research and development tool rather than a technique ready for an automated implementation for the routine verification of forecasts.

Acknowledgements

NCEP and ECMWF ensemble forecast data reported in the paper are archived in the TIGGE dataset (http://apps.ecmwf.int/datasets/data/tigge/). Stage IV analyses are available from https://data.eol.ucar.edu/dataset/21.093. This research has been conducted as part of the NOAA MAPP S2S Prediction Task Force and supported by NOAA grant NA16OAR4311082. Finally, we would like to thank the anonymous reviewers for their helpful comments.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by Climate Program Office [grant number NA16OAR4311082].

Notes

1. While the practice of naming winter storms is controversial, the collection of named storms provides a representative sample of precipitation events that a major player of the US weather enterprise expected to have potentially high impact on society.
2. For instance, for an idealized precipitation field with a regular shape, $\boldsymbol{r}$ could be defined by the centre of mass of the precipitation field. It should be noted, however, that for a complex precipitation field, the estimate of the location error by our technique is not necessarily equal to the error in the location of the centre of mass (Han and Szunyogh, 2016).

References

Bengtsson, L. and Körnich, H. 2016. Impact of a stochastic parametrization of cumulus convection, using cellular automata, in a mesoscale ensemble prediction system. *Q. J. Roy. Met. Soc.* **142**(695), 1150–1159.

Buizza, R. 1997. Potential forecast skill of ensemble prediction and spread and skill distribution of the ECMWF ensemble prediction system. *Mon. Wea. Rev.* **125**, 99–119.

Greybush, S. J., Saslo, S. and Grumm, R. 2017. Assessing the ensemble predictability of precipitation forecasts for the January 2015 and 2016 east coast winter storms. *Wea. Forecasting* **32**(3), 1057–1078.

Han, F. and Szunyogh, I. 2016. A morphing-based technique for the verification of precipitation forecasts. *Mon. Wea. Rev.* **144**, 295–313.

Han, F. and Szunyogh, I. 2017. A technique for the verification of precipitation forecasts and its application to a problem of predictability. Submitted to Mon. Wea. Rev. Preprint Online at: https://sites.google.com/a/tamu.edu/hanfan/.

Herrera, M. A., Szunyogh, I. and Tribbia, J. 2016. Forecast uncertainty dynamics in the Thorpex Interactive Grand Global Ensemble (TIGGE). *Mon. Wea. Rev.* **144**(7), 2739–2766.

Keil, C. and Craig, G. C. 2007. A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Wea. Rev.* **135**, 3248–3259.

Keil, C. and Craig, G. C. 2009. A displacement and amplitude score employing an optical flow technique. *Wea. Forecasting* **24**, 1297–1308.

Khouider, B., Biello, J., Majda, A. J., et al. 2010. A stochastic multicloud model for tropical convection. *Commun. Math. Sci.* **8**(1), 187–216.

Lin, Y. and Mitchell K. E. 2005. The NCEP Stage II/IV hourly precipitation analyses: development and applications. In: *19th Conference of Hydrology*, American Meteorological Society.

Loeser, C. F., Herrera, M. A. and Szunyogh, I. 2017. An assessment of the performance of the operational global ensemble forecast systems in predicting the forecast uncertainty. *Wea. Forecasting* **32**(1), 149–164.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., and co-authors 2009. Stochastic parametrization and model uncertainty. *ECMWF Tech. Memo* **598**, 1–42.

Talagrand, O., Vautard R. and Strauss B. 1999. Evaluation of probabilistic prediction systems. In: *Proceedings of the ECMWF Workshop on Predictability*, European Centre for Medium Range Weather Forecasts, Shinfield Park, Reading, Berkshire, UK, 1–25.

Wang, Z. and Bovik, A. C. 2002. A universal image quality index. *IEEE Signal Process. Lett.* **9**(3), 81–84. DOI:10.1109/97.995823.

Wang, Z. and Bovik, A. C. 2009. Mean squared error: love it or leave it? A new look at signal fidelity measures. *IEEE Signal Process. Mag.* **26**, 98–117.

Wang, Z., Bovik, A. C., Sheikh, H. R. and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612.

Wernli, H., Paulat, M., Hagen, M. and Frei, C. 2008. SAL – a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Wea. Rev.* **136**, 4470–4487.