Measure of Forecast Challenge and Predictability Horizon Diagram Index for Ensemble Models

JUN DU

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

BINBIN ZHOU

MSG at NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

JASON LEVIT

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

(Manuscript received 11 July 2018, in final form 20 December 2018)

ABSTRACT

Responding to the call for new verification methods in a recent editorial in *Weather and Forecasting*, this study proposed two new verification metrics to quantify the forecast challenges that a user faces in decision-making when using ensemble models. The measure of forecast challenge (MFC) combines forecast error and uncertainty information together into one single score. It consists of four elements: ensemble mean error, spread, nonlinearity, and outliers. The cross correlation among the four elements indicates that each element contains independent information. The relative contribution of each element to the MFC is analyzed by calculating the correlation between each element and MFC. The biggest contributor is the ensemble mean error, followed by the ensemble spread, nonlinearity, and outliers. By applying MFC to the predictability horizon diagram of a forecast ensemble, a predictability horizon diagram index (PHDX) is defined to quantify how the ensemble evolves at a specific location as an event approaches. The value of PHDX varies between 1.0 and -1.0 . A positive PHDX indicates that the forecast challenge decreases as an event nears (type I), providing creditable forecast information to users. A negative PHDX value indicates that the forecast challenge increases as an event nears (type II), providing misleading information to users. A near-zero PHDX value indicates that the forecast challenge remains large as an event nears, providing largely uncertain information to users. Unlike current verification metrics that verify at a particular point in time, PHDX verifies a forecasting process through many forecasting cycles. Forecasting-process-oriented verification could be a new direction in model verification. The sample ensemble forecasts used in this study are produced from the NCEP global and regional ensembles.

1. Introduction

Given that ensemble forecasts are able to not only provide a forecast but also the uncertainty or confidence associated with the forecast, ensemble models have become an increasingly important part of numerical weather prediction (Buizza et al. 2018; Du et al. 2018). A forecast that does not provide quantitative uncertainty information is incomplete (National

Research Council 2006). Currently, to verify an ensemble of forecasts, forecast error and uncertainty are measured by separate metrics (Du and Zhou 2017). These separate metrics are useful in helping model developers pinpoint the causes of forecast problems for model improvement. For example, the ensemble mean error might be related to a model deficiency, while the problem of spread could be caused by the ensemble perturbation methods. However, users of weather forecasts need to consider both the expected forecast outcome and the forecast confidence at the same time in their decision-making process. Forecasts with the same ensemble mean (i.e., same error) but different spreads (Fig. 1) pose a very different level of

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

Corresponding author: Dr. Jun Du, jun.du@noaa.gov

DOI: 10.1175/WAF-D-18-0114.1

For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](http://www.ametsoc.org/PUBSReuseLicenses).

FIG. 1. Two forecast ensembles (members in black and ensemble mean in red) with similar ensemble means but different spreads. Ensemble forecast 2 is more difficult to use than ensemble 1 in decision-making.

challenges for decision-making since this could lead to different actions based on the cost–loss ratio (Du and Deng 2010). Therefore, it is necessary and useful to have a single combined verification metric to address both forecast error and uncertainty information from a user’s point of view. Such a new metric, the measure of forecast challenge (MFC), is proposed in section 2c.

Another major challenge that a decision-maker faces is the time evolution of a forecast over different forecast cycles. For example, forecast confidence increases as an event nears in some cases, while that is not so in other cases. Apparently, making a decision is relatively easier for the former, while it is extremely difficult for the latter. Some forecasts become more accurate while others become less accurate as an event approaches, which provides a very different quality of information to a user. Unfortunately, this important piece of information about the forecast time evolution is missing in current verification metrics. Current verification metrics do not include forecasting process information but merely a snapshot of a particular forecast at a selected time. To quantitatively measure the “forecasting process” of a forecast ensemble, a predictability horizon diagram index (PHDX) is proposed in section 2d, by applying the MFC to a specific event at a specific location. Finally, a summary is given in section 3.

2. Data and method

a. Data

The ensemble forecasts used in this study are produced by the NCEP Global Ensemble Forecast System (GEFS; Zhou et al. 2017) and the Short-Range Ensemble Forecast (SREF; Du et al. 2015). The GEFS has 21 members at approximately 34-km horizontal resolution. The base model of GEFS is the NCEP Global Forecast System (EMC 2018b). The SREF has 26 members at 16-km horizontal resolution with two regional models: the Nonhydrostatic Multiscale Model on the B Grid (NMMB; EMC 2018a) and ARW (MMM 2018). The Climatology Calibrated Precipitation Analysis (CCPA; Hou et al. 2014) is used as truth for precipitation, while the GFS analysis (Kleist and Ide 2015) is used as truth for the other variables.

b. Existing metrics

Four elements can be derived from an ensemble of forecasts to exhibit different aspects of the forecast challenge. They are the ensemble mean forecast error, spread, nonlinearity, and outliers. Ensemble mean forecast error (EME) measures the average skill of a model; for example, EME is the absolute error of the ensemble mean forecast m compared to an observation o :

FIG. 2. The four ensemble-derived elements related to predictability: (a) absolute error of ensemble-mean forecast, (b) ensemble spread, (c) nonlinearity, and (d) relative outlier for 850-hPa geopotential height at 126-h lead time, initialized at 0000 UTC 30 Dec 2017. The background contour is the control forecast.

$$\text{EME} = |m - o|. \quad (1)$$

Ensemble spread (Sprd) is related to forecast uncertainty and is defined as the standard deviation of n ensemble members m_i with respect to their ensemble mean m :

$$\text{Sprd} = \sqrt{\frac{1}{n} \times \sum_{i=1}^n (m_i - m)^2}, \quad (i = 1, 2, 3, \dots, n). \quad (2)$$

The absolute difference between an ensemble control run m_{ctl} and ensemble mean forecast m is a measure of the nonlinearity (NonLN) of atmospheric flow (Du and Zhou 2011):

$$\text{NonLN} = |m - m_{\text{ctl}}|. \quad (3)$$

An ensemble outlier (OUT) examines if the truth falls outside of an ensemble envelope (Du and Zhou 2017). Quantitatively, OUT measures how far an observation falls outside of an ensemble envelope, defined in this study as a “relative outlier,” that is, the ratio of the

outside distance (between the observation and the nearest edge of the envelope or “absolute outlier”) to the width of the ensemble envelope (the distance between the maximum and minimum members m_{max} and m_{min}):

$$\text{OUT} = \frac{o - m_{\text{max}}}{m_{\text{max}} - m_{\text{min}}}, \quad \text{if } o > m_{\text{max}}, \quad (4a)$$

$$\text{OUT} = \frac{m_{\text{min}} - o}{m_{\text{max}} - m_{\text{min}}}, \quad \text{if } o < m_{\text{min}}, \quad \text{and} \quad (4b)$$

$$\text{OUT} = 0, \quad \text{if } m_{\text{min}} \leq o \leq m_{\text{max}}. \quad (4c)$$

Figure 2 shows these four elements for 850-hPa geopotential height over North America for a Northeast coastal winter storm case (Kocin and Uccellini 2004), which occurred along the U.S. East Coast around 4 January 2018. These elements have similarities and differences in spatial structures. To quantify the similarities and differences, spatial correlations (Pearson correlation) of the six exhaustive unique pairs among the four elements were calculated for four representative

FIG. 3. Spatial correlation coefficients (%) of the six exhaustive unique pairs out of the four ensemble-derived elements [EME, spread (Spr), nonlinearity (NL), and outlier (Out)] over the global domain for (a) H850, (b) 850-hPa temperature (T850), (c) RH850, and (d) 300-hPa meridional wind (V300), varying with forecast lead time (h). The GEFS is initialized at 0000 UTC 30 Dec 2017.

variables (geopotential height, temperature, and moisture at the 850-hPa level, and wind at the 300-hPa level) over the global domain (point by point over 259 920 grid points, only the North American region is displayed in Fig. 2 as well as later, in Fig. 4) and shown in Fig. 3. The highest correlation is found between the spread and nonlinearity (~ 0.6 or 60% at 126 h), followed by $\sim 45\%$ between the ensemble mean error and spread. The correlation is about 35% between the ensemble mean error and outlier, and about 30% between ensemble mean error and nonlinearity. There is a slightly negative correlation for the remaining two pairs: “nonlinearity and outlier” (-10%) and “spread and outlier” (-20%). The correlation varies slightly with variable. Therefore, each of these four elements has its own unique information due to their difference from each other in their spatial structure, with the collinearity existing to some degree among them.

c. Measure of forecast challenge

We incorporate these four ensemble-derived elements defined by Eqs. (1)–(4) into a combined score called the measure of forecast challenge by using Eq. (5). This is intended to measure the forecast challenge a user faces when he or she uses the ensemble forecasts to make a decision:

$$\text{MFC} = (\text{EME} + \text{Sprd} + \text{NonLN}) \times (1 + \text{OUT}). \quad (5)$$

The meteorological assumption behind Eq. (5) is that if the ensemble-mean forecast error is larger, the ensemble spread is larger, and the nonlinearity is stronger, it is a more challenging forecast to use in decision-making. The score is then further punished by OUT if an observation falls outside of the ensemble envelope. Therefore, MFC reflects not only the ensemble mean error but also the forecast uncertainty, the nature of the flow (linear

FIG. 4. MFC of H850, T850, RH850, and V300 over the North America domain at 126-h lead time. The background contour is the control forecast. The GEFS is initialized at 0000 UTC 30 Dec 2017.

or nonlinear), and the systematic deficiencies (biases) in the model, initial conditions, and ensemble perturbation techniques.

Figure 4 shows the spatial distribution of MFC for the four variables over the North American domain. To examine the relative contributions of the four elements in shaping the MFC, the spatial correlation between MFC (Fig. 4) and each of the four elements [as in Fig. 2 for the 850-hPa geopotential height (H850)] is calculated over the global domain and shown in Fig. 5. Figure 5 shows that the contribution remained almost constant over all forecast hours for the ensemble mean error, increased with forecast time for the ensemble spread and nonlinearity, and decreased for the outliers [except for 850-hPa relative humidity (RH850)]. At the 126-h forecast length, the most significant contributor is the ensemble mean error ($\sim 85\%$ in correlation), followed by the ensemble spread ($\sim 70\%$), nonlinearity ($\sim 60\%$), and outlier ($\sim 40\%$). Although the ensemble mean error is the most significant contributor to MFC, contributions from the other elements cannot be neglected. For example, by comparing Figs. 4a and 2a, the

modification to the ensemble mean error's contribution by the other elements is obviously noticeable. At local scales (e.g., at a grid point), these modifications to the ensemble mean error's contribution could be impactful (see section 2d).

d. Application of MFC to the predictability horizon diagram

The time evolution of a forecast over multiple forecast cycles is an important piece of information for a user in making a confident decision in the real world. To see how an ensemble of forecasts evolves at a specific location as an event approaches, a “predictability horizon diagram” was used by Greybush et al. (2017). The diagram displays the ensemble forecast from different cycles, hence at different lead times, verified at a common valid time ($t = 0$). Since an event normally becomes more and more predictable as its time nears, one expects to see the evolution of “detection” (from no members to a few members capturing the event), “emergence of signal” (more and more members surrounding the event but with a larger spread), and

FIG. 5. Spatial correlation coefficients (%) between the MFC and each component (EME, Spr, NL, and Out) over the global domain for (a) H850, (b) T850, (c) RH850, and (d) V300, varying with forecast lead time (h). The GEFS is initialized at 0000 UTC 30 Dec 2017.

“convergence of solutions” (a majority or all of the members surrounding the event with decreasing spread) from an ensemble of forecasts as the forecast lead time shortens (Fig. 6a). To evaluate the challenge levels of different forecasts with respect to users, a quantitative score is desired to measure the process of this time evolution. Therefore, the PHDX is defined in this study by applying the MFC to the diagram. Figure 6b shows the four elements used to calculate the MFC, and Fig. 6c is the resulting MFC for this ideal case over forecast lead time ($t = 12, 11, \dots, 2, 1$). We can see that the MFC decreases as forecast lead time decreases in this situation (type I). Figure 7a demonstrates the opposite situation: some ensemble members did capture the event in the longer-range forecasts but all members drifted away from the truth in the shorter-range forecasts. The corresponding MFC

(Fig. 7c) indicates that it becomes more and more challenging as the forecast lead time becomes shorter (type II). The situation in Figs. 8a and 9a lies between those of Figs. 6a and 7a, where the ensemble does not converge to a solution but remains diverse as the event time nears. In Fig. 8a the ensemble members are scattered around the observation, while in Fig. 9a the members are all located below the observation (a “negative” model bias). As expected, the distribution of MFC (Figs. 8c and 9c) indicates similar forecast challenges over forecast lead time (type III) for both cases while the magnitude of MFC indicates that Fig. 8 is less challenging than Fig. 9.

Based on the above analysis of Figs. 6–9, a predictability horizon diagram index has been designed to quantitatively summarize these characteristics. Conceptually, PHDX is proportional to the net trend of

FIG. 6. An ideal schematic case of a type-I situation for an event (obs). (a) Evolution of ensemble forecasts over different forecast cycles (lead time) during three stages: detection, emergence of signal, and convergence of solutions. (b) The four ensemble-derived elements: EME, Sprd, NonLN, and OUT. (c) MFC.

MFC over forecast lead time (decreasing, increasing, or fluctuating) and inversely proportional to the magnitude of MFC:

$$\text{PHDX} = \frac{\text{Trend}}{\text{Mag}}, \quad (6)$$

where Trend is the net trend over the forecast cycle or lead time t ($t = 12, 11, \dots, 2, 1$ in Figs. 6–9), defined as

$$\text{Trend} = \sum_T^2 (\delta_t \times \text{Avslp}), \quad (t = T, T-1, \dots, 3, 2). \quad (7)$$

Here, T is the oldest cycle, and 1 is the most current cycle. The average slope or change between two neighboring forecast cycles is Avslp and is defined as

$$\text{Avslp} = \frac{1}{T-1} \times \sum_T^2 |\text{MFC}(t-1) - \text{MFC}(t)|, \quad (t = T, T-1, \dots, 3, 2), \quad (8)$$

and where

$$\delta_t = 1, \quad \text{if } \text{MFC}(t-1) - \text{MFC}(t) < 0 \quad (\text{decreasing of MFC}), \quad (9a)$$

$$\delta_t = -1, \quad \text{if } \text{MFC}(t-1) - \text{MFC}(t) > 0 \quad (\text{increasing of MFC}), \quad \text{and} \quad (9b)$$

$$\delta_t = 0, \quad \text{if } \text{MFC}(t-1) - \text{MFC}(t) = 0 \quad (\text{no change of MFC}). \quad (9c)$$

In addition, Mag is the total magnitude of MFC over all forecast cycles and is defined as

$$\text{Mag} = \sum_T^1 \text{MFC}(t). \quad (10)$$

The PHDX varies from -1.0 to 1.0 , where a positive value is associated with a type-I situation, a negative value with a type-II situation, and a near-zero value with a type-III situation. A user receives creditable information from an ensemble of forecasts in type I, misleading information in type II, and largely uncertain information in type III to use in making a decision.

FIG. 7. As in Fig. 6, but for a type-II situation.

The PHDX values for Figs. 6–9 are 0.30, -0.11 , 0.00, and -0.01 , respectively.

For real-world cases, the predictability horizon diagram pattern could be complicated for some variables, such as precipitation. Figures 10a–c are for precipitation forecasts from real cases with day-to-day weather predicted by the NCEP SREF, representing the type-I (PHDX = 0.10), type-II (-0.17), and type-III (0.02) forecasting processes, respectively. For other variables such as 2-m temperature, the time evolution more closely follows the type-I situation (Fig. 10d), and the forecasting process typically provides more creditable information (PHDX = 0.12) compared to precipitation. This different behavior between precipitation and temperature suggests that the proposed new scores for MFC and PHDX can easily identify less predictable variables from more predictable variables since precipitation is indeed less predictable than temperature in general. By the way, at a station site the modification to the ensemble mean error's contribution by the other elements could be significant. For example, as shown in Fig. 10c, the time evolution of the ensemble mean error shows that it increased from 87 to 75 h and then decreased from 75 to 63 h, but the MFC actually behaved just the

opposite: it decreased from 87 to 75 h and increased from 75 to 63 h.

The PHDX is also applied to two major winter storms (28 January 2015 and 25 January 2016). The predictability of the heavy precipitation associated with these two storms has been thoroughly studied by Greybush et al. (2017) using NCEP GEFS data. Their Figs. 4 and 5 showed the predictability horizon diagrams of the storm-total precipitation forecasts for the two storms, respectively, at three major metropolitan areas (Boston, Massachusetts; New York, New York; and Washington, D.C.). Based on their predictability horizon diagrams, we calculated the PHDX values to determine which location received the more useful forecast information. For the 28 January 2015 storm, the PHDX values are 0.08 for Boston and 0.0 for New York. This is consistent with their analysis: the ensemble solutions converged toward the observation as the event approached (type I) for Boston (their Fig. 4a), while they never converged, with large uncertainty (type III), for New York (their Fig. 4b). Similarly for the 25 January 2016 storm, the PHDX is 0.07 for Washington (their Fig. 5a) and 0.02 for New York (their Fig. 5b). This is again consistent with their analysis: the GEFS provided credible information

FIG. 8. As in Fig. 6, but for a type-III situation, where the ensemble encompasses the event.

to Washington area forecasters and largely uncertain information to New York forecasters.

3. Summary

Two new verification metrics are proposed to measure forecast challenges that a user faces in making a decision when employing ensemble forecasts. Unlike some verification approaches that focus on the model diagnosis (e.g., Han and Szunyogh 2018), these two new metrics are not intended for diagnosing forecast problems but for assessing the challenge levels a user faces in his or her real-world application. One is the measure of forecast challenge, which combines forecast error and uncertainty information together into one single score (analogically like evolving a 2D photo into a 3D photo). Specifically, MFC consists of the four predictability-related elements derived from an ensemble: ensemble mean error, spread, nonlinearity, and outliers. Using NCEP GEFS data, we examined the cross correlation among the four elements for four representative variables (850-hPa geopotential height, 850-hPa temperature, 850-hPa relative humidity, and 300-hPa meridional wind). Moderate-to-low correlation was found for them. For example, on average at the

126-h forecast length, the highest correlation is found between the spread and nonlinearity ($\sim 60\%$), followed by $\sim 45\%$ between the ensemble mean error and spread. The correlation is about 35% between the ensemble mean error and outlier, and about 30% between the ensemble mean error and nonlinearity. There is a slightly negative correlation for the remaining two pairs: the nonlinearity and outlier (-10%) and the spread and outlier (-20%). The correlation varies slightly with variable. Therefore, these four elements have their own unique information to contribute in shaping the MFC.

The relative contribution of each component to MFC is then analyzed by calculating the correlation between each component and MFC for the same four variables. The biggest contributor is the ensemble mean error, which remained almost constant over all forecast hours ($\sim 85\%$ in correlation). The second biggest contributor is the ensemble spread, which increased with forecast time ($\sim 70\%$ at 126 h). The contribution of nonlinearity also increased with forecast time and reached $\sim 60\%$ in correlation at 126 h. Since the outlier acts as a local amplifier [Eq. (5)], it is the smallest contributor. The outlier's contribution decreased with forecast time and was about 30% in correlation at 126 h. Although

FIG. 9. As in Fig. 6, but for a type-III situation, where the ensemble has systematic bias and largely misses the event.

the ensemble mean error is the most significant contributor to MFC, the contributions from the other elements cannot be neglected. The modification to the ensemble mean error's contribution by the other elements is visibly noticeable at both domain and local scales.

By applying MFC to the predictability horizon diagram of a forecast ensemble, another new score: the predictability horizon diagram index (PHDX) has been proposed. It is designed to quantify how the ensemble evolves through multiple forecast cycles at a specific location as an event approaches (analogically like evolving a photo into a video by adding a time dimension). The value of PHDX varies between 1.0 to -1.0 and can be used to compare ensemble performances from a user's point of view. It can be categorized into three types. A positive PHDX is associated with the situation where the forecast challenge decreases as an event nears (type I). Type I provides credible forecast information to users. In contrast, a negative PHDX value indicates a situation where the forecast challenge increases as an event nears (type II). Type II provides misleading information to users. A type-III situation lies between types I and II, where the PHDX value is near zero and the forecast

challenge remains large even as the forecast length shortens. Type III provides largely uncertain information to users. Using NCEP SREF data as well as two major winter storm cases (derived from another researcher's study), we demonstrated that the PHDX can successfully quantify the difficulty level of a forecasting process for operational ensemble models. We can also use PHDX to compare different fields; for example, precipitation is generally more difficult to forecast than surface temperature. By the way, it will be interesting to investigate further how the distribution of a variable impacts the behavior of the scores. For example, a precipitation forecast is often highly skewed in its distribution while temperature is normally in a Gaussian distribution; metrics like ensemble mean error and spread may be not as informative for precipitation as for temperature. By grouping precipitation forecasts into different categories (i.e., heavy rain, light rain), will the MFC behave more stably for each category of precipitation forecasts?

Finally, we want to emphasize that unlike current verification metrics, which act like a snapshot of a forecast at a particular point in time, PHDX verifies a forecasting process through multiple forecast cycles. Forecasting-process-oriented verification could be a new direction in

FIG. 10. As in Figs. 6–9, but for real day-to-day cases from the NCEP SREF. Precipitation forecasts with (a) type-I and (b) type-II forecasting processes.

FIG. 10c. As in (a) and (b) but for the type-III forecasting process. (d) The 2-m temperature forecast.

verifying NWP models. We encourage researchers and forecasters to use this novel forecasting-process-based approach to verify ensemble models.

Acknowledgments. This is a part of the authors' regular work at EMC/NCEP/NOAA. Ms. Mary Hart of NCEP is appreciated for improving the readability of our manuscript. An internal review was done by Ying Lin and Perry Shafran. We also thank the three anonymous reviewers and the editor, Brian Ancell, for their constructive suggestions that greatly improved the presentation of this work.

REFERENCES

- Buizza, R., J. Du, Z. Toth, and D. Hou, 2018: Major operational ensemble prediction systems (EPS) and the future of EPS. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, https://doi.org/10.1007/978-3-642-40457-3_14-1.
- Du, J., and G. Deng, 2010: The utility of the transition from deterministic to probabilistic weather forecasts—Verification and application of probabilistic forecasts. *Meteor. Mon.*, **36**, 10–18.
- , and B. Zhou, 2011: A dynamical performance-ranking method for predicting individual ensemble member performance and its application to ensemble averaging. *Mon. Wea. Rev.*, **139**, 3284–3303, <https://doi.org/10.1175/MWR-D-10-05007.1>.
- , and —, 2017: Ensemble fog prediction. *Marine Fog: Challenges and Advancements in Observations, Modeling, and Forecasting*, D. Koracin and C. E. Dorman, Eds., Springer, 477–509, https://doi.org/10.1007/978-3-319-45229-6_10.
- , G. DiMego, B. Zhou, D. Jovic, B. Ferrier, and B. Yang, 2015: Short Range Ensemble Forecast (SREF) system at NCEP: Recent development and future transition. *23rd Conf. on Numerical Weather Prediction/27th Conf. on Weather Analysis and Forecasting*, Chicago, IL, Amer. Meteor. Soc., 2A.5, <https://ams.confex.com/ams/27WAF23NWP/webprogram/Paper273421.html>.
- , and Coauthors, 2018: Ensemble methods for meteorological predictions. *Handbook of Hydrometeorological Ensemble Forecasting*, Q. Duan et al., Eds., Springer, https://doi.org/10.1007/978-3-642-40457-3_13-1.
- EMC, 2018a: NOAA Environmental Modeling System. NWS/Environmental Modeling Center, <http://www.emc.ncep.noaa.gov/NEMS/doc.php>.
- , 2018b: The Global Forecast System (GFS)—Global Spectral Model (GSM) (GSM version 13.0.2). NWS/Environmental Modeling Center, <http://www.emc.ncep.noaa.gov/GFS/doc.php>.
- Greybush, S., S. Saslo, and R. Grumm, 2017: Assessing the ensemble predictability of precipitation forecasts for the January 2015 and 2016 East Coast winter storms. *Wea. Forecasting*, **32**, 1057–1078, <https://doi.org/10.1175/WAF-D-16-0153.1>.
- Han, F., and I. Szunyogh, 2018: How well can an ensemble predict the uncertainty in the location of winter storm precipitation? *Tellus*, **70A**, 1440870, <https://doi.org/10.1080/16000870.2018.1440870>.
- Hou, D., and Coauthors, 2014: Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of Stage IV toward CPC gauge-based analysis. *J. Hydrometeor.*, **15**, 2542–2557, <https://doi.org/10.1175/JHM-D-11-0140.1>.
- Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational–ensemble data assimilation for the NCEP GFS. Part I: System description and 3D-hybrid results. *Mon. Wea. Rev.*, **143**, 433–451, <https://doi.org/10.1175/MWR-D-13-00351.1>.
- Kocin, P. J., and L. W. Uccellini, 2004: *Northeast Snowstorms*. Vols. 1 and 2, *Meteor. Monogr.*, No. 54, Amer. Meteor. Soc., 818 pp.
- MMM, 2018: NCAR Weather Research and Forecasting Model (Advanced Research WRF). Mesoscale and Microscale Meteorology Laboratory, NCAR–UCAR, <https://www.mmm.ucar.edu/weather-research-and-forecasting-model>.
- National Research Council, 2006: Completing the forecast – Characterizing and communicating uncertainty for better decisions using weather and climate forecasts. NRC Rep., The National Academies Press, 112 pp.
- Zhou, X., Y. Zhu, D. Hou, Y. Lou, J. Peng, and R. Wobus, 2017: Performances of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, <https://doi.org/10.1175/WAF-D-17-0023.1>.