Mesoscale Ensemble Weather Prediction at U.S. Army Dugway Proving Ground, Utah

JASON C. KNIEVEL,[a] YUBAO LIU,[a] THOMAS M. HOPSON,[a] JUSTIN S. SHAW,[a] SCOTT F. HALVORSON,[b] HENRY H. FISHER,[a] GREGORY ROUX,[a] RONG-SHYANG SHEU,[a] LINLIN PAN,[a] WANLI WU,[a,d] JOSHUA P. HACKER,[a,e] ERIK VERNON,[b] FRANK W. GALLAGHER III,[c] AND JOHN C. PACE[b]

[a] *National Center for Atmospheric Research,*[f] *Boulder, Colorado*
[b] *U.S. Army Dugway Proving Ground, Dugway, Utah*
[c] *National Oceanic and Atmospheric Administration, Silver Spring, Maryland*

(Manuscript received 20 April 2017, in final form 21 September 2017)

ABSTRACT

Since 2007, meteorologists of the U.S. Army Test and Evaluation Command (ATEC) at Dugway Proving Ground (DPG), Utah, have relied on a mesoscale ensemble prediction system (EPS) known as the Ensemble Four-Dimensional Weather System (E-4DWX). This article describes E-4DWX and the innovative way in which it is calibrated, how it performs, why it was developed, and how meteorologists at DPG use it. E-4DWX has 30 operational members, each configured to produce forecasts of 48 h every 6 h on a 272-processor high performance computer (HPC) at DPG. The ensemble's members differ from one another in initial-, lateral-, and lower-boundary conditions; in methods of data assimilation; and in physical parameterizations. The predictive core of all members is the Advanced Research core of the Weather Research and Forecasting (WRF) Model. Numerical predictions of the most useful near-surface variables are dynamically calibrated through algorithms that combine logistic regression and quantile regression, generating statistically realistic probabilistic depictions of the atmosphere's future state at DPG's observing sites. Army meteorologists view E-4DWX's output via customized figures posted to a restricted website. Some of these figures summarize collective results—for example, through means, standard deviations, or fractions of the ensemble exceeding thresholds. Other figures show each forecast, individually or grouped—for example, through spaghetti diagrams and time series. This article presents examples of each type of figure.

1. Introduction

a. Background

In operational forecasting, ensemble prediction systems (EPSs) have become a mainstay for addressing two unavoidable truths: 1) numerical weather prediction (NWP) models will always be flawed, and 2) the chaotic atmosphere's true state will always elude any attempt at perfect observations. Every individual simulation from an NWP model is compromised by flaws in the model's numerical schemes, physical parameterizations, and methods of data assimilation; by imperfections in initial conditions, boundary conditions, and assimilated observations; by limitations in computers and networks; and by dynamical instabilities in the atmosphere itself (inertial, convective, baroclinic, etc.). No *individual* simulation can capture the uncertainties that arise from these flaws, imperfections, and limitations. An *ensemble* of simulations can (Toth et al. 2001).

EPSs offer other advantages, too. The mean prediction by an ensemble tends to be more skillful than any single prediction by one of its members (Leith 1974; Du et al. 1997; Toth and Kalnay 1997; Ebert 2001; Ma et al. 2012). The spread among members' predictions can be interpreted roughly to indicate the mean prediction's uncertainty (Kalnay 2003). Ensembles also permit

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

[d] Current affiliation: Spire, Boulder, Colorado.
[e] Current affiliation: Jupiter Technology Systems, Boulder, Colorado.
[f] The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author: Dr. Jason Knievel, knievel@ucar.edu

DOI: 10.1175/WAF-D-17-0049.1

output from NWP systems to be framed as probabilities. Probabilistic guidance from an EPS is potentially much more useful to decision-makers than is traditional guidance from a single NWP model (e.g., Buizza 2008).

Designing an EPS for predicting mesoscale weather is challenging. Over the finer two subranges of the mesoscale (horizontal distances of 2–200 km), weather phenomena are smaller, more transitory, and less predictable than the phenomena for which many early EPSs were designed (Hamill et al. 2000; Hohenegger and Schär 2007). Initial states do not necessarily comprise the full range of resolvable scales, so mesoscale NWP models must generate finescale phenomena and processes during forward integration, and often need finescale perturbations to be prescribed so uncertainty in the analysis is properly represented (e.g., Toth and Kalnay 1993; Johnson et al. 2014; Iyer et al. 2016). At smaller grid intervals, complex interactions occur over a wider range of scales, which affects an ensemble's spread and errors (Hohenegger and Schär 2007; Clark et al. 2009; Eckel et al. 2010; Johnson et al. 2014).

Most mesoscale models are limited-area models, so lateral boundary conditions heavily influence solutions (McDonald 1997; Pielke 2002), and uncertainty in boundary conditions is generally underrepresented (Nutter et al. 2004). Unlike global EPSs, which tend to emphasize differences in the initial conditions (e.g., Molteni et al. 1996), mesoscale EPSs tend to emphasize uncertainties in models—for example, treatments of physical processes such as radiative transfer, formation of clouds and precipitation, and eddies in the boundary layer (Bouttier et al. 2012). Many methods for generating ensemble perturbations are less appropriate for short-term modeling at the fine mesoscale than for modeling at the temporal and spatial scales for which the methods were originally developed (Eckel and Mass 2005). Sufficient natural spread can be elusive (Hamill et al. 2000; Eckel et al. 2010; Romine et al. 2014; Schwartz et al. 2014). Large errors in mesoscale models that are not fully taken into account can translate to unrealistically small ensemble spreads and large systemic errors in an ensemble as a whole (Eckel et al. 2010; Berner et al. 2015).

Despite such challenges to ensemble prediction with mesoscale models, phenomena of mesoscale size and duration nonetheless are very important for many users of numerical weather predictions. Meteorologists at the U.S. Army Dugway Proving Ground (DPG), Utah are one such group. Since 2007, they have used an EPS known as the Ensemble Four-Dimensional Weather System (E-4DWX) as a primary tool for supporting test exercises (Liu et al. 2007). In 2014, E-4DWX was extended to three more government test sites in the Great Basin of the United States: White Sands Missile Range (WSMR), New Mexico; Yuma Proving Ground (YPG), Arizona; and Electronic Proving Ground (EPG), Arizona. For brevity, this paper focuses on just DPG's experience with E-4DWX.

b. *Testing and forecasting at DPG*

One of DPG's primary missions is to test equipment that detects chemical and biological hazards. Such tests are very sensitive to mesoscale and microscale weather. Numerical predictions influence whether and how tests are conducted. Numerical analyses influence how the results from tests are interpreted. Skillful guidance is required about the dispersion of chemical or biological agent simulants that are released into the open air near the ground. This guidance must be based on detailed and accurate numerical predictions of temperature, static stability, wind speed, and wind direction over mesoscale and microscale distances and times.

It can be costly to delay or cancel a test because of poor weather guidance. The nature of testing at DPG and other sites of the U.S. Army Test and Evaluation Command (ATEC), and the weather-related decisions that have to be made there by users of NWP forecasts, are particularly amenable to probabilistic guidance, which can feed into cost–loss analyses and similar frameworks in ways that deterministic guidance cannot. For example, if you know the cost C and loss L associated with taking

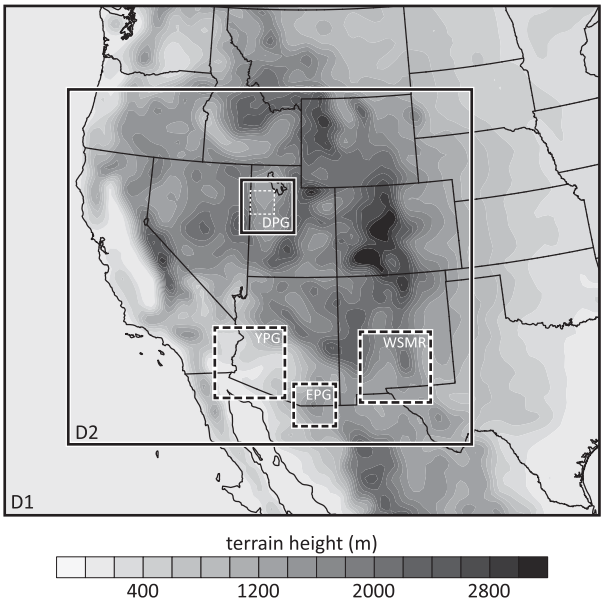

FIG. 1. Domains 1 (dx = 30 km), 2 (dx = 10 km), and 3 (dx = 3.3 km) of E-4DWX. Only DPG has a formal domain 3. For the three other test facilities (YPG, EPG, and WSMR) local subregions of domain 2 are used for the ensemble products. The thin dashed white line inside DPG's domain 3 marks the region depicted in Fig. 2.

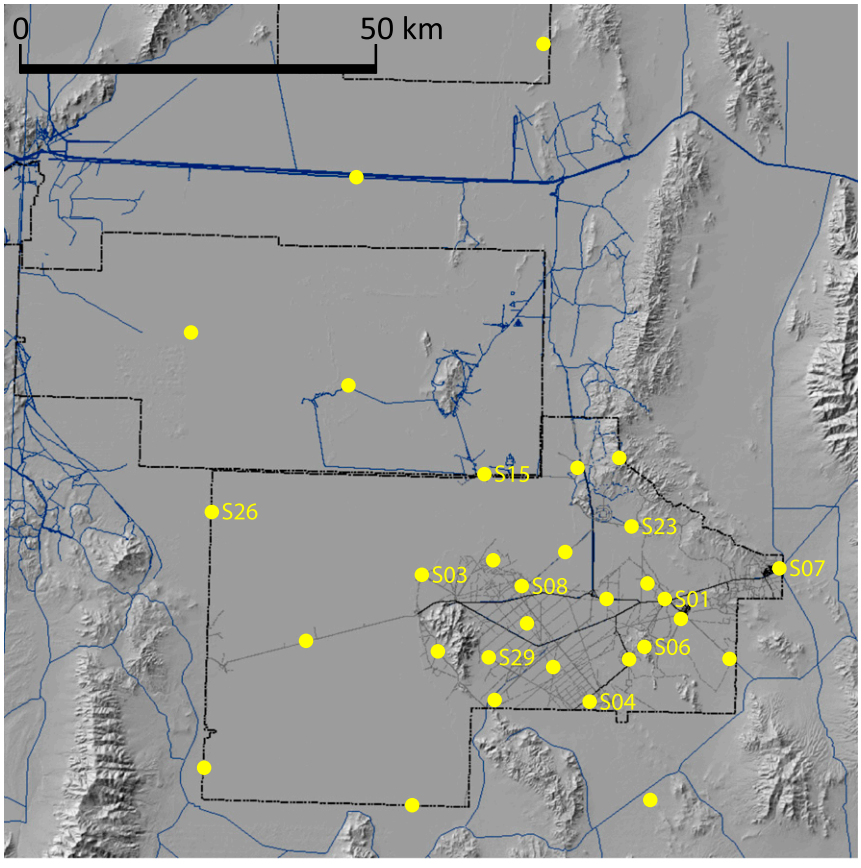

FIG. 2. Locations of SAMS observing stations at DPG. Numbered stations are the sources of data that are the basis for subsequent figures. To provide geographic context, the perimeter of this region is outlined in Fig. 1.

action (e.g., canceling a test) to avoid certain weather (e.g., 10-m wind speed $> 20\,\mathrm{m\,s^{-1}}$), then you should take action if $P > C/L$, wherein P is the probability of that weather as predicted by a calibrated ensemble (Murphy 1977). (There is much more about calibration in section 2d.)

c. *Weather at DPG*

Semiarid DPG is located at the southern end of the Great Salt Lake Desert, roughly 100 km southwest of Salt Lake City, Utah, on the bed of ancient Lake Bonneville. Narrow, compact mountain ranges rise more than a kilometer above the dry soil and scrubby vegetation on the desert floor. During high pressure weather patterns, nocturnal drainage flows often develop along the slopes of the mountains and the gently inclined terrain at their bases, and the nocturnal boundary layer is often strongly stable (e.g., Rife et al. 2002; Lehner et al. 2015; Grachev et al. 2016; Jeglum et al. 2017). Vast, smooth salt flats, or playas, lay to the west and north of the primary test sites at DPG (Massey et al. 2017). Playas are moist and highly reflective, so the near-surface air above them is relatively cool during the day and relatively warm at night. The resultant local temperature gradients drive thermally direct "salt breezes" (Rife et al. 2002). Gap flows accelerate through low spots in the ranges under certain conditions (Jeglum et al. 2017).

Interacting with these and other locally forced phenomena are additional synoptic, mesoscale, and microscale features that characterize weather in the Intermountain West (Fernando et al. 2015; Jeglum and Hoch 2016). These complex interactions challenge meteorologists at DPG and the E-4DWX NWP system that they use (described in section 2 below). Years of monitoring how members contribute to the overall skill of the ensemble have demonstrated that no single configuration of E-4DWX outperforms all others in every case, and no single source of initial and boundary conditions has proved altogether superior. Therefore, rather than run a single model with a configuration selected a priori, it is more effective to run an ensemble of configurations that perform well in at least some circumstances.

E-4DWX was first installed at DPG in 2007. Since then, we are not aware of a single test at DPG that was canceled solely because of unpredicted adverse weather.

TABLE 1. The E-4DWX members at the time of writing. Please see the technical note on the WRF Model v.3 (Skamarock et al. 2008) for more information about the schemes in the table. Options for LBCs are NAM and GFS. The option for the land surface model (LSM) is Noah. Options for the surface-layer scheme (Surf) are the MM5 Monin–Obukhov (M–O), Eta, Mellor–Yamada–Nakanishi–Niino (MYNN), and quasi-normal scale elimination (QNSE). Options for the ABL scheme are Yonsei University (YSU), Mellor–Yamada–Janjić (MYJ), Bougeault–Lacarrère (B–L), MYNN level 2.5, QNSE, and University of Washington (aka Bretherton and Park) (UW). Options for the cumulus scheme (Cu), which is only applied on domains 1 and 2 except as noted, are Kain–Fritsch (K–F), Grell–Freitas (G–F), and Betts–Miller–Janjić (BMJ). Options for the microphysics scheme (Micro) are WRF single-moment 6-class (WSM6), new Thompson et al. (Th), WRF single-moment 5-class (WSM5), Morrison double-moment (Mor), and WRF double-moment 5-class (WDM5). Options for the shortwave radiation scheme (SW) are Goddard (Gd), Dudhia (Du), and Community Atmosphere Model (CAM). Options for the longwave radiation scheme (LW) are Rapid Radiative Transfer Model (RRTM) and CAM. The last column denotes the two control runs (control), a configuration with a cumulus scheme used on domain 3 (Cu D3), a configuration without data assimilation (No DA), two configurations to which stochastic kinetic-energy backscatter is applied (SKEBS) with different random number streams (1 and 2), and two configurations that have lower boundary conditions shifted by 30 km (Shift) and employ a Kalman filter (KF) with different nudging coefficients (1 and 2).

Member	LBC	LSM	Surf	ABL	Cu	Micro	SW	LW	Notes
1	NAM	Noah	M–O	YSU	K–F	WSM6	Gd	RRTM	Control
2	GFS	Noah	M–O	YSU	K–F	WSM6	Gd	RRTM	Control
3	NAM	Noah	M–O	YSU	G–F	WSM6	Gd	RRTM	
4	NAM	Noah	Eta	MYJ	K–F	WSM6	Gd	RRTM	
5	NAM	Noah	M–O	YSU	K–F	Th	Gd	RRTM	
6	NAM	Noah	M–O	YSU	K–F	WSM5	Gd	RRTM	
7	NAM	Noah	M–O	YSU	K–F	Mor	Gd	RRTM	
8	GFS	Noah	Eta	B–L	K–F	WSM6	Gd	RRTM	
9	GFS	Noah	MYNN	MYNN	K–F	WSM6	Gd	RRTM	
10	GFS	Noah	QNSE	QNSE	K–F	WSM6	Gd	RRTM	
11	GFS	Noah	M–O	UW	K–F	WSM6	Gd	RRTM	
12	GFS	Noah	M–O	YSU	K–F	WSM6	CAM	CAM	
13	GFS	Noah	M–O	YSU	BMJ	WSM6	Gd	RRTM	
14	GFS	Noah	M–O	YSU	K–F	WSM6	Du	RRTM	
15	GFS	Noah	M–O	YSU	G–F	WSM6	Gd	RRTM	
16	GFS	Noah	M–O	YSU	BMJ	WSM6	Gd	RRTM	Cu D3
17	NAM	Noah	Eta	B–L	K–F	WSM6	Gd	RRTM	
18	NAM	Noah	QNSE	QNSE	K–F	WSM6	Gd	RRTM	
19	NAM	Noah	MYNN	MYNN	K–F	WSM6	Gd	RRTM	
20	NAM	Noah	M–O	YSU	K–F	WSM6	Du	RRTM	
21	NAM	Noah	M–O	YSU	K–F	WSM6	CAM	CAM	
22	NAM	Noah	M–O	YSU	BMJ	WSM6	Gd	RRTM	
23	GFS	Noah	M–O	YSU	K–F	Th	Gd	RRTM	
24	GFS	Noah	M–O	YSU	K–F	WDM6	Gd	RRTM	
25	GFS	Noah	M–O	YSU	K–F	Mor	Gd	RRTM	
26	GFS	Noah	M–O	YSU	K–F	WSM6	Gd	RRTM	No DA
27	GFS	Noah	M–O	YSU	K–F	WSM6	Gd	RRTM	SKEBS 1
28	GFS	Noah	M–O	YSU	K–F	WSM6	Gd	RRTM	SKEBS 2
29	GFS	Noah	M–O	YSU	K–F	WSM6	Du	RRTM	Shift KF 1
30	GFS	Noah	M–O	YSU	K–F	WSM6	Du	RRTM	Shift KF 2

Meteorologists attribute some of their success to E-4DWX. There have been situations in which E-4DWX predicted that marginal weather was likely, and a test was conducted anyway because the risk was deemed acceptable. In these situations, test participants can sometimes accomplish a reduced set of objectives. There have also been situations in which enough ensemble members predicted unacceptable weather that a test was canceled, even though the standard deterministic forecast that is run as baseline guidance at DPG predicted acceptable weather.

2. E-4DWX

a. Framework and NWP core

E-4DWX is a generic, multitier framework that integrates processing and assimilation of observations with ensemble prediction, can be rapidly configured for use anywhere over the globe, and can readily incorporate new advancements by the community engaged in research on mesoscale ensemble prediction.

A skillful EPS relies on a skillful deterministic NWP model and a high quality data assimilation system

TABLE 2. Regressors used to calibrate selected variables from E-4DWX.

Regressor	Notes	Used in
24-h persistence	Observed weather 24 h earlier	LR, QR
Ensemble median	Ensemble-median forecast at validation time	LR, QR
Ensemble mean	Ensemble-mean forecast at validation time	LR, QR
Ensemble spread	Standard deviation of the forecast ensemble at validation time	LR, QR
Ensemble mean divided by the forecast standard deviation	Steepness of logistic function can vary with standard deviation of the forecast ensemble	LR, QR
Constant	Climatology	LR, QR
LR-predicted quantile	Retained from LR step in calibration	QR
Corresponding quantile	Corresponding (interpolated) quantile from the raw 30-member forecast ensemble	QR
Each member of the 30-member (uncalibrated) raw ensemble	Each member of the raw 30-member forecast ensemble used as independent regressor	LR, QR

(Buizza et al. 2005). An EPS built on a poor model will produce forecast errors that are rooted in the deficiencies of the model itself, not in the inevitable uncertainties in the atmosphere's initial state (Kalnay 2003). The NWP core of E-4DWX is the Advanced Research version of the Weather Research and Forecasting (WRF) Model (v.3.5.1 at the time of writing). Early versions of E-4DWX also used the fifth-generation Mesoscale Model (MM5) developed by The Pennsylvania State University and NCAR, but the current version does not. The framework of E-4DWX is sufficiently flexible that in the future a variety of other mesoscale models could also be added.

E-4DWX comprises 30 operational NWP members (i.e., excluding test members). Each runs on three one-way nested computational domains of grid intervals dx of 30.0, 10.0, and 3.3 km (Fig. 1). There are 37 levels, roughly one-third of which are in the lowest 1 km. The single 3.3-km domain of 76 × 76 points is positioned over DPG. Forecast data for the other three supported test facilities are drawn from the larger, coarser domain with the 10-km grid interval. The 48-h forecasts are initialized four times per day at 0000, 0600, 1200, and 1800 UTC. E-4DWX is executed at DPG on a 272-processor Linux cluster.

Atmospheric observations, some from special observing platforms known as surface atmospheric measurement systems (SAMS; Fig. 2), are assimilated by E-4DWX via the Real-Time Four-Dimensional Data Assimilation System (RTFDDA; Liu et al. 2008), which employs Newtonian relaxation to nudge the model toward observations during the assimilation cycle. This is a computationally efficient, robust method of assimilating data continuously rather than intermittently, thereby pairing more closely the time of an observation to its corresponding time in the simulation (Stauffer and Seaman 1994).

Part of the spread among members in E-4DWX develops from differences in the initial conditions (ICs), lateral boundary conditions (LBCs), and land surface (LS) characteristics. Some members draw their ICs and LBCs from the North American Mesoscale Forecast System (NAM), others from the Global Forecast System (GFS). To approximate the effects of phase errors in LBCs from the NAM and GFS, the fields from the two models are shifted horizontally by 30 km (one grid cell) west–east and north–south for two members. This method and the distance of 30 km are based on our trial-and-error tests; shifts of two grid cells produce unrealistically large errors. To perturb ICs for some members, we modify the assimilated observations and the weights and radii of influence in RTFDDA. The modifications to the observations are in the form of random and fixed errors (biases) up to 1.5°C and 1.5 m s^{-1}, chosen empirically to be consistent with RTFDDA's typical analysis errors near the ground and in the lower troposphere. Finally, physical parameterizations vary among most members, so E-4DWX includes multimodel spread as well. Table 1 summarizes the 30 configurations.

To make greatest use of E-4DWX, meteorologists at DPG need the probabilistic predictions from the system to comprise realistic heterogeneity on the temporal and spatial scales of the weather that most influences tests. There are many challenges to using mesoscale ensembles for true probabilistic predictions spanning from several hours to several days. In addition to the challenges described in the introduction, NWP models are often biased. One way of mitigating bias (among other undesirable qualities) in an EPS is through calibration (Warner 2011). That is the focus of the next subsection.

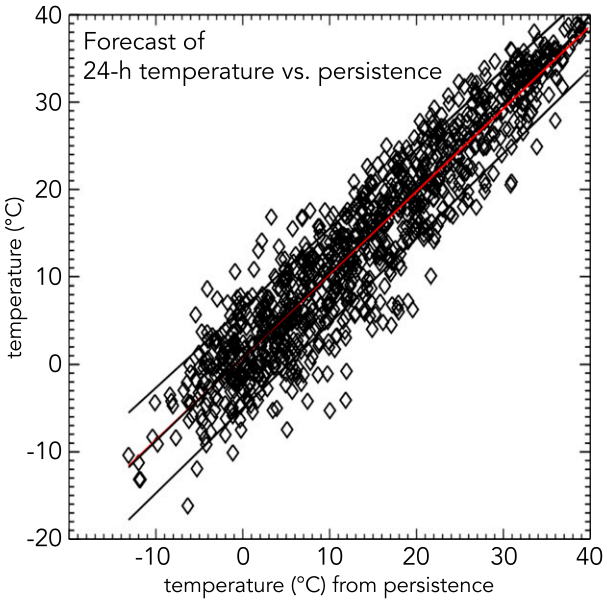

FIG. 3. Air temperature at 2 m (AGL in all figures) at DPG's SAMS 01 during June–August 1979–2001. Each value from the previous day is used as a 24-h persistence forecast. The red line is the fit of the central tendency (mean) from the standard linear regression, the middle black line is the fit of the 0.5 quantile (median), the upper black line is the 0.9 quantile, and the lower black line is the 0.1 quantile. The mean and median are similar but diverge as temperature increases. The fits of the 0.1 and 0.9 quantiles converge as temperature increases, indicating the data are heteroscedastic.

b. Calibration

During postprocessing, part of the output from E-4DWX is calibrated—made statistically reliable—such that the probability of conditions predicted by E-4DWX matches as well as possible the observed frequency of those conditions (Wilks 2006b). This means that E-4DWX's predictions are statistically indistinguishable from observations for the subset of variables being calibrated: surface air pressure, air temperature at 2 m (AGL throughout the article), relative humidity at 2 m, and wind speed and vector wind components at 10 m. Currently, calibration is being applied only to predictions interpolated to instrumented sites. Benefits of calibration include the following:

- reducing the forecast error of the ensemble mean, partly by reducing bias; the mean of a calibrated, properly perturbed, and sufficiently large ensemble theoretically has on average as little as one-half the error variance of any of the ensemble's members (Leith 1974);
- increasing the ensemble's reliability, resolution, and sharpness, including for predicting the likelihood of extreme and potentially devastating weather (e.g., Hamill et al. 2004)—calibration does not *guarantee* more skillful predictions of extremes, however (e.g., Mylne 2002); and
- providing an indication of forecast uncertainty through the spread among ensemble members (e.g., Hagedorn et al. 2012), an indication that often is limited and inexact (e.g., Hopson 2014).

Over the last several decades, methods for calibrating EPSs and for evaluating their calibration have been based on linear regression (e.g., Atger 2003; Diomede et al. 2014), logistic regression (e.g., Hamill et al. 2004; Wilks 2006a; Hamill et al. 2008; Bentzien and Friederichs 2012; Johnson and Wang 2012; Roulin and Vannitsem 2012), nonhomogeneous Gaussian regression (e.g., Gneiting et al. 2005; Hagedorn et al. 2008), ensemble kernel density model output statistics (e.g., Glahn et al. 2009), ensemble dressing (e.g., Roulston and Smith 2003; Wang and Bishop 2005; Fortin et al. 2006; Wilks and Hamill 2007), ensemble regression (e.g., Unger et al. 2009), Bayesian model averaging (e.g., Raftery et al. 2005; Sloughter et al. 2007; Wilson et al. 2007; Kleiber et al. 2011), spatial Bayesian model averaging (e.g., Berrocal et al. 2007), simultaneous quantile regression (e.g., Tokdar and Kadane 2012), quantile-to-quantile mapping (e.g., Hamill and Whitaker 2006; Diomede et al. 2014), rank histograms and other indicators of reliability (e.g., Hamill and Colucci 1997; Eckel and Walters 1998; Krzysztofowicz and Sigrest 1999; Atger 2003; Johnson and Wang 2012), spread–skill relationships (e.g., Atger 1999), analogs (e.g., Hamill and Whitaker 2006; Diomede et al. 2014; Junk et al. 2015), geostatistical model averaging (e.g., Kleiber et al. 2011), parametric mixture models (e.g., Bentzien and Friederichs 2012), object-based methods (e.g., Nehrkorn et al. 2014), and artificial neural networks (e.g., Yuan et al. 2007).

The calibration developed for E-4DWX is novel in two ways. 1) We combine logistic regression (Hamill and Whitaker 2006; Wilks and Hamill 2007) with quantile regression (Koenker and Bassett 1978) to improve numerical predictions at discrete, evolving probability intervals rather than at fixed climatological thresholds. 2) To ensure the ensemble's reliability (Hopson 2014), we preprocess it, then explicitly condition the training of the final postprocessing model on the (now calibrated) ensemble dispersion. All aspects of the calibration algorithms employ a common framework to ensure that the results are statistically robust. Regressions are always performed with cross-validation to minimize the likelihood of overfitting; in each pass through the dataset, half of the data are used to fit the regression, and the other half are used to evaluate the fit. The subsamples are then reversed. Logistic regression and quantile regression are explained in more detail in the next two subsections.

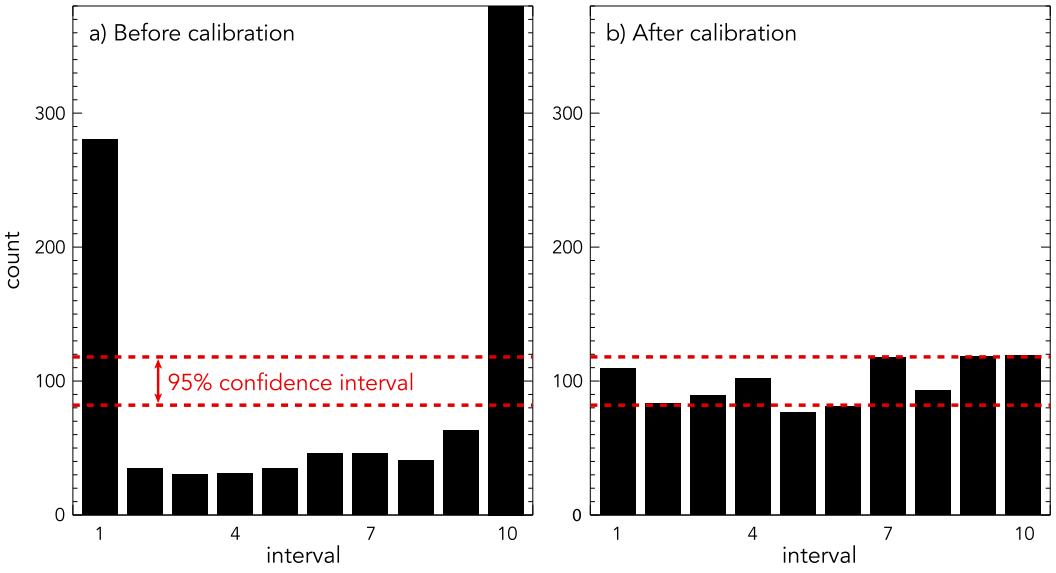

FIG. 4. Rank histograms of E-4DWX's (a) uncalibrated and (b) calibrated 24-h forecasts of 2-m air temperature at DPG's SAMS 01 every 6 h at 1331 different valid times from 0000 UTC 8 Jun 2013 through 1800 UTC 14 Jul 2014. The red dashed lines bound the 95% confidence limits for perfectly calibrated forecasts. Both histograms are based on nine quantiles that separate the data into 10 intervals (i.e., bins).

1) LOGISTIC REGRESSION

Logistic regression (LR) is a well-established approach for fitting data to the logit function (e.g., Hilliker and Fritsch 1999; Hamill and Whitaker 2006; Wilks and Hamill 2007). It is useful for our application because probabilistic predictions—for instance, from an ensemble such as E-4DWX—can be expressed categorically as meeting or not meeting a series of thresholds. Logistic regression handles such cases of binary predictands quite well. It results in probabilities that are correctly bounded by 0 and 1, and it accommodates residuals that are non-Gaussian. Please see the text by Wilks (2006b) for more information about logistic regression, and for a comparison between it and alternative methods of regression for binary predictands. Papers by Messner et al. (2014a,b) offer examples of recent advancements in LR.

2) QUANTILE REGRESSION

When a probability density function (PDF) is segmented into two or more probability intervals, a quantile is what separates one interval from another. For example, if a PDF were divided into two such intervals of equal probability, the single quantile between them would be the median. Quantile regression (QR) is an absolute-error estimator that can conditionally fit specific quantiles of a regressand's distribution (beyond just the median) without relying on the assumption that the regressand or residuals are distributed parametrically (Koenker and Bassett 1978). Probability distributions need not be Gaussian, for example. QR offers several other benefits, too. Because the conditional fit in QR is based on absolute error, it is less sensitive to outliers than are squared-error estimators (Portnoy and Koenker 1997). QR also accommodates distributions of data with heteroscedastic variances (i.e., when variance is a function of a predictand's magnitude). One of the first applications of QR in atmospheric science was by Bremnes (2004) to forecasts of precipitation.

The QR algorithms used to fit specific quantiles can be explained as follows. Assume $\{y_i\}$ is a set of observations of the regressand y, and $\{\mathbf{x}_i\}$ is an associated set of predictors. Just as in standard linear regression, a linear function of $\mathbf{x}$ can be used to estimate a specific quantile q_θ of y:

$$q_\theta(\mathbf{x}_i;\boldsymbol{\beta}) = \beta_0 + \sum_{k=1}^{n} \beta_k x_{ik} + r_i, \qquad (1)$$

with residual $r_i = y_i - q_\theta(\mathbf{x}_i;\boldsymbol{\beta})$and $\theta \in (0, 1)$, wherein $\boldsymbol{\beta}$ is a vector of unknown coefficients. However, instead of minimizing the squared residuals, as is done with standard linear regression, in QR a weighted iterative minimization of $\{r_i\}$ is performed to estimate $\boldsymbol{\beta}$:

$$\min \sum_{i=1}^{n} \rho_\theta(r_i) = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \rho_\theta[y_i - q_\theta(\mathbf{x}_i;\boldsymbol{\beta})], \qquad (2)$$

with a weighting function of arbitrary quantity a defined as

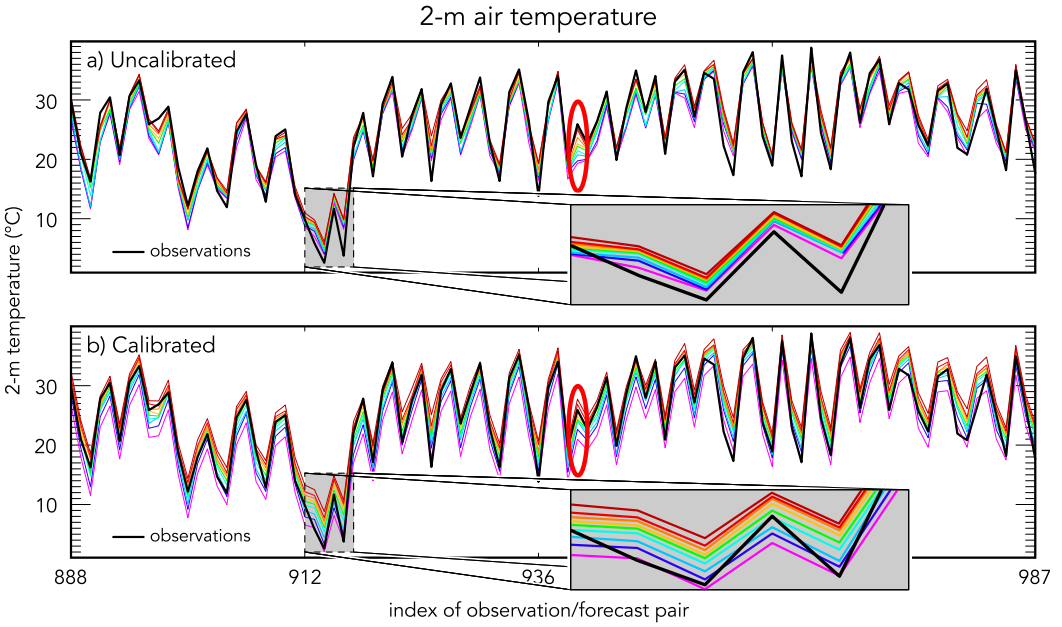

FIG. 5. Time series of (a) uncalibrated and (b) calibrated 24-h forecasts (color) from E-4DWX at 100 different times from 0000 UTC 11 Jun 2014 (time 888) through 1200 UTC 13 Jul 2014 (time 987). Superposed on the panels is the observed 2-m air temperature (°C; black) every 6 h at DPG's SAMS 01. Calibration mitigates periods of bias in the uncalibrated forecast and increases the ensemble spread, as enlarged for clarity in the gray inset boxes, which span the same arbitrarily chosen period of time. The red oval marks 26 Jun 2014, the valid time of the data presented in Fig. 6. Data in these two panels are ordered sequentially at intervals of 6 h for times when data were available; spans of missing data are skipped.

$$\rho_\theta(a) = \begin{cases} \theta a, & a \geq 0 \\ (\theta - 1)a, & a < 0 \end{cases}. \quad (3)$$

A powerful benefit of QR is that minimizing the cost function in (2) leads to a statistically flat rank histogram, which characterizes a calibrated ensemble prediction (i.e., a prediction that is equivalent to a random draw from an underlying—usually unknown—distribution). In addition, although QR constrains the resultant quantile estimators to satisfying this requirement, at the same time it also constrains the estimators to optimal sharpness (Wilks 2006b), generating PDFs that are narrower than a PDF from a purely climatological distribution.

3) SPECIFIC STEPS FOR COMBINING LR AND QR

In the first step of our implementation of LR and QR to calibrate E-4DWX, we start with an archive of observations at the sites for which we want calibrated guidance. From that archive we estimate the climatological PDF of a variable of interest, such as 2-m air temperature. This is the *prior* PDF. To characterize this distribution, 99 equally spaced quantiles $q_{c,m}$ are used to segment the PDF into 100 bins. For each quantile,

TABLE 3. MAE in 10-m wind speed ($m\,s^{-1}$) and 2-m temperature (°C) at SAMS 01, 04, and 08 in uncalibrated and calibrated 24-h forecasts from E-4DWX at 100 different times from 0000 UTC 11 Jun through 1200 UTC 13 Jul 2014. The time span matches the time series in Fig. 5.

		MAE ($m\,s^{-1}$ or °C)	
Station	Variable	Before calibration	After calibration
SAMS 01	10-m wind speed	1.20	1.08
SAMS 04	10-m wind speed	1.26	1.19
SAMS 08	10-m wind speed	1.20	1.13
SAMS 01	2-m temperature	2.18	2.00
SAMS 04	2-m temperature	1.80	1.73
SAMS 08	2-m temperature	1.97	1.83

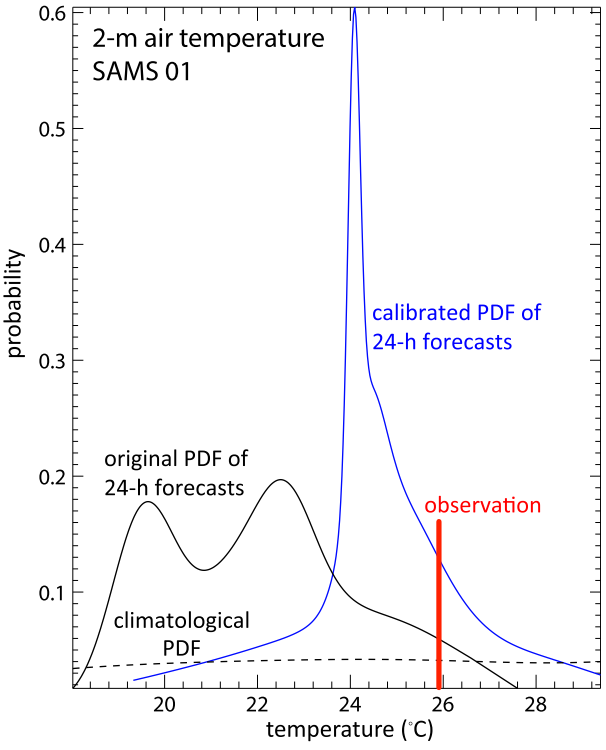

FIG. 6. Example of PDFs of an original uncalibrated (solid black) and a calibrated (blue) 24-h ensemble forecast from E-4DWX of 2-m air temperature (°C) at DPG's SAMS 01 at 1800 UTC 26 Jun 2014. Comparison to the observation of 25.9°C (red bar) and the climatological PDF for summer (dashed black) demonstrates how calibration reduces bias, broadens spread, and increases sharpness.

$$\Pr(V \le q_{c,m}) = \frac{m}{M+1}, \tag{4}$$

wherein V is the validation value, such as from a future observation, and m is one of the M (99, in this case) quantiles that separate the equally spaced probability bins to be calibrated.

Using past numerical predictions and their corresponding observations—these two datasets together compose the *training data*—we then apply LR to estimate out-of-sample models for exceeding each of the 99 quantiles. For each prediction day, a conditional cumulative distribution function (CDF) is formed. The predictors' values (input) are properties of an individual forecast. If the past numerical predictions used for training are skillful, the result is a conditional CDF that is sharper than one from climatology. At this stage in the process, the conditional CDF from the training data is still defined in terms of the 99 quantiles.

The next step is to interpolate those 99 quantiles of the conditional CDF to fewer N quantiles. For E-4DWX, we chose $N = 9$. The specific choice is arbitrary. It should balance computational cost and granularity, which $N = 9$ does for our purposes. Granularity is particularly important if one is concerned about precision at the ends of the distribution, where the least likely and most likely outcomes are found. With $N = 9$, the lowest probability threshold that can be predicted is 10%, the highest 90%. Values that have probabilities in the range of 0%–10% are all grouped into the former bin, 90%–100% into the latter. To discriminate between probabilities of 1% and 2%, for instance, or between 98% and 99%, one would need at least $N = 99$.

As follows from the definition of quantile presented above, a calibrated N-member ensemble will have $N + 1$ equally weighted probability bins, and for each we can find corresponding predictions of variables simply by choosing daily predicted quantiles $q_{n,f}$ associated with probability $\Pr(V \le q_{n,f}) = n/(N+1)$, wherein n is an ensemble member from an (interpolated) ensemble of size N. The resultant nine-member (i.e., 9 quantiles, 10 bins) interpolated ensemble's PDF is narrower than the PDF from the original 99 quantiles, as one would expect. We call this ensemble *interpolated* to emphasize that it is not the group of original, unmodified 30 dynamical members of E-4DWX, but rather an "ensemble" of nine quantiles that describe the statistical distribution of the training data.

Next, QR is applied to the nine quantiles that now characterize the training data, following the equations presented in the previous subsection. The regressors for the QR include the LR quantiles from the earlier step, along with the other regressors listed in Table 2. A graphical depiction of how QR optimally determines the relationship of a regressor on specific quantiles of the regressand, with no parametric assumptions, is presented in Fig. 3. To make the example clear and simple, the data in Fig. 3 are from a single regressor: 24-h persistence forecasts of 2-m air temperature at SAMS 01 at DPG (Fig. 2) during June–August 1979–2001, to which (1)–(3) have been applied for the 0.1, 0.5 (median), and 0.9 quantiles. The red line is the fit for the central tendency (mean) calculated from standard linear regression. The black lines, from bottom to top, are the fits for $q_{0.1}$, $q_{0.5}$ (median), and $q_{0.9}$. The fits for the mean and median are similar, but they diverge with increasing temperature. That the black lines fitted to the three quantiles are not parallel signifies a heteroscedastic distribution.

Applying LR and QR using regressors that are functions of ensemble spread (Table 2) accounts for day-to-day variability in the spread and for varying dynamical stability (error growth rate). Nearly flat rank histograms are also guaranteed by a good fit of the quantile regression equations. But regressors based on ensemble spread do not guarantee *resolution* in the ensemble

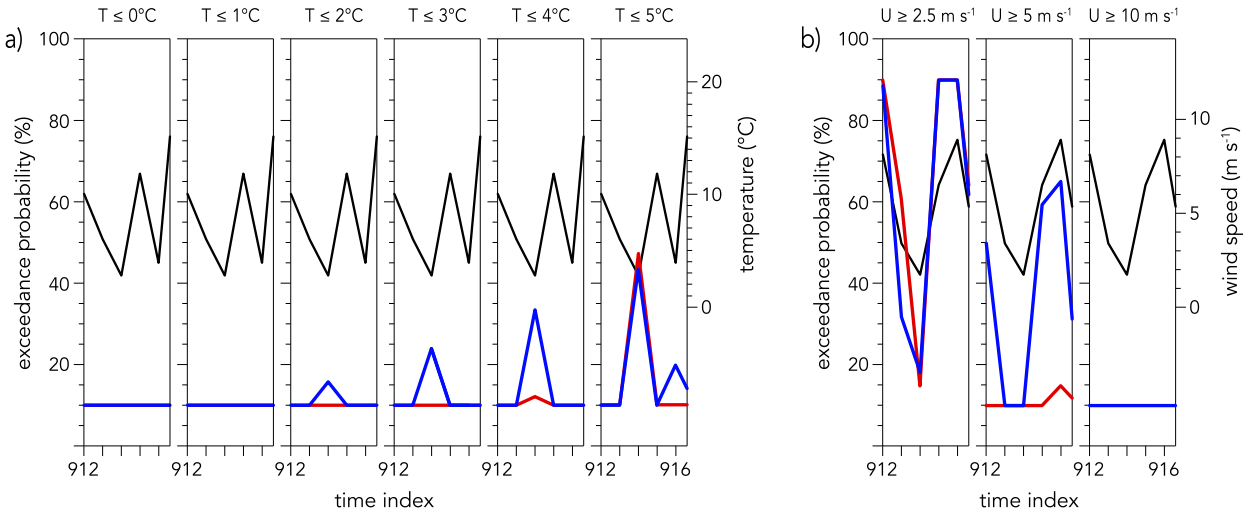

FIG. 7. E-4DWX's uncalibrated (red) and calibrated (blue) probability (%) of (a) 2-m air temperature ≤ 0°–5°C (from left to right) and (b) 10-m wind speed ≥ 2.5, 5.0, and 10.0 m s^{-1} (from left to right) at SAMS 01 for the span of time depicted in the gray inset boxes in Fig. 5. For reference, observed 2-m air temperature and 10-m wind speed are also plotted (black) in (a) and (b), respectively. Uncalibrated and calibrated probabilities are constrained to a minimum of 10% and a maximum of 90% owing to the choice of nine quantiles [see section 2d(3)]. Probabilities within the range 10%–90% are interpolated and therefore are not limited to multiples of 10%.

predictions. For example, consider the pathological case pointed out by Hamill (2001) in which a flat rank histogram can belie an ensemble that might not sample from the correct distribution of any single observation used for validation. This can occur if LR and QR are both fitted to a dataset that spans many dynamical regimes.

To overcome this pitfall, in the next step we explicitly condition the training data on the ensemble dispersion. We sort the predicted nine daily forecast quantiles (from the QR step above) into three equally populated bins of low, medium, and high dispersion, which ensures that the predictions are reliable with respect to the full dataset. The choice of three bins for the sorting is not requisite, but the number is limited by the training sample size. Sorting into bins can be thought of as identifying "spread analogs" in the local state space. For other examples of analogs in ensemble postprocessing, please see the papers by Hamill and Whitaker (2006), Hopson and Webster (2010), and Delle Monache et al. (2013).

After the predictions are sorted, the QR equations are then applied anew to the distributions in each of the three bins (i.e., on the conditioned dataset), using the same procedure described above. This allows the raw ensemble predictions to self-diagnose periods of relative stability or instability, which informs its own postprocessing.

In the final step of processing the training data, the results from the initial application of the QR—*before* sorting data by dispersion—are compared to the results from QR applied *after* sorting data by dispersion. For each site and forecast lead time, we determine which is the better of the two applications of QR, and it is *that* QR model that is applied to calibrate the current forecast for a given site and forecast lead time. The metrics we use for determining the better QR model are Brier score, rank probability skill score, root-mean-square error (RMSE), receiver operating characteristic (ROC), and a measure of the rank histogram's flatness.

Once every step of processing the training data is completed (as described above), the calibration of the current ensemble prediction consists of submitting the output from the ensemble members to a very similar series of steps, starting with the LR and ending with the QR model chosen for the given site and lead time, as described in the previous paragraph.

3. Performance

a. *Effectiveness of calibration*

Rank histograms of forecasts from 9 June 2013 through 13 July 2014 are evidence that calibration produces realistically dispersive, unbiased ensemble forecasts (Fig. 4). The U-shaped histogram before calibration (Fig. 4a) signifies that too many observations fall outside of the envelope of members' predictions (i.e., insufficient dispersion). The histogram's asymmetry, heavier to the right, signifies that the number of observations higher than the highest prediction exceeds the number of observations lower than the lowest prediction (i.e., low bias). Calibration corrects these two

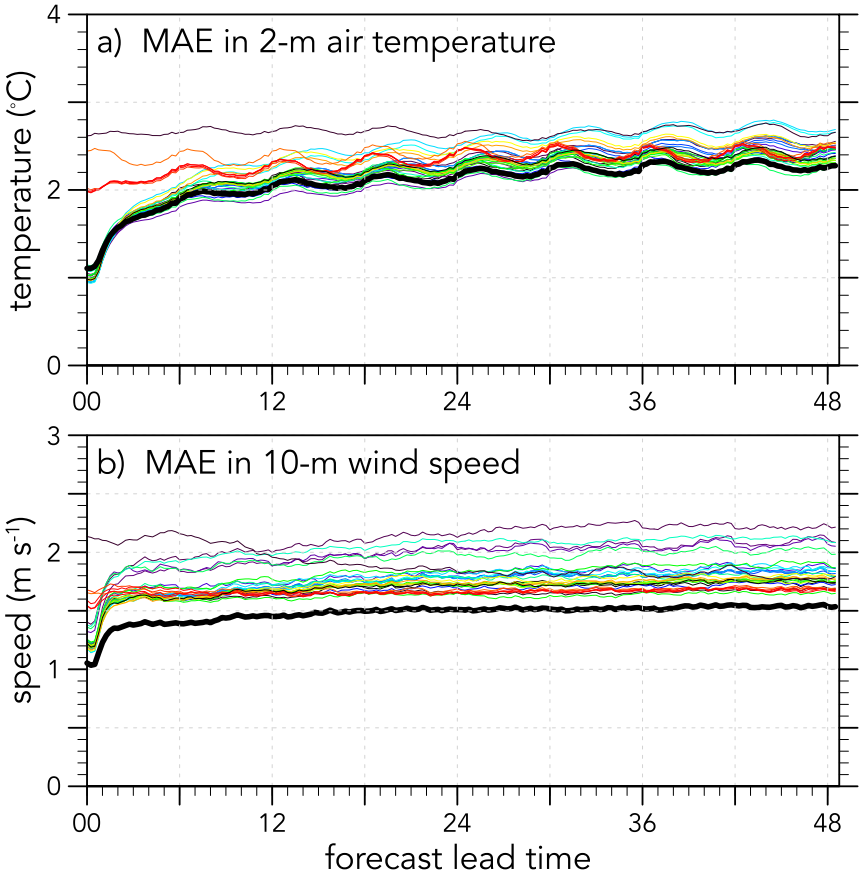

FIG. 8. MAE in (a) 2-m air temperature (°C) and (b) 10-m wind speed ($m\,s^{-1}$) as a function of forecast lead time on E-4DWX's domain 3 (dx = 3.3 km) in 2015 calculated from uncalibrated forecasts interpolated to 31 observation sites. The thick black line marks the ensemble mean. The thin colored lines mark individual ensemble members. Member 26 in purple (see Table 1) has the largest initial MAE and approximately flat error growth thereafter because it has no data assimilation.

deficiencies (Fig. 4b). Results are similar at other lead times and for other variables (not shown).

How calibration modifies output from E-4DWX's members at individual valid times is more clearly illustrated in Fig. 5, which presents uncalibrated and calibrated time series of 24-h predictions of 2-m air temperature at DPG's SAMS 01 (Fig. 2). The modifications are easiest to see in the gray inset boxes that enlarge the same short, arbitrarily chosen window of time on the panels. Calibration broadens the envelope of temperatures spanned by the members and recenters the midpoint between the members with the lowest and highest forecasts. As a result, the envelope more frequently bounds the time series of observations that for verification were superposed on the forecasts. Also as a result, mean forecasts are improved (Table 3). The degree of improvement depends, of course, on the skill of the uncalibrated mean, which is already quite high for the stations listed in Table 3. (The next subsection elaborates on the skill of the uncalibrated E-4DWX.)

Drawn from the time series in Fig. 5, the individual case of 26 June 2014 (Fig. 6) further illustrates the improvements in the first and second moments of the ensemble PDF. The biased, uncalibrated ensemble (solid black line) has a bimodal PDF mostly to the left of the observation (red bar). The unbiased, calibrated PDF (blue line) is a more useful forecast because it is reliable and sharp. The climatological PDF (dashed black line) is reliable but not sharp. The uncalibrated ensemble is sharper than climatology but less reliable.

Where the uncalibrated PDF in Fig. 6 is located along the x axis, and where the envelope of members' predictions is located along the y axis in Fig. 5, are functions of the uncalibrated E-4DWX's bias. However, bias cannot be determined from any individual

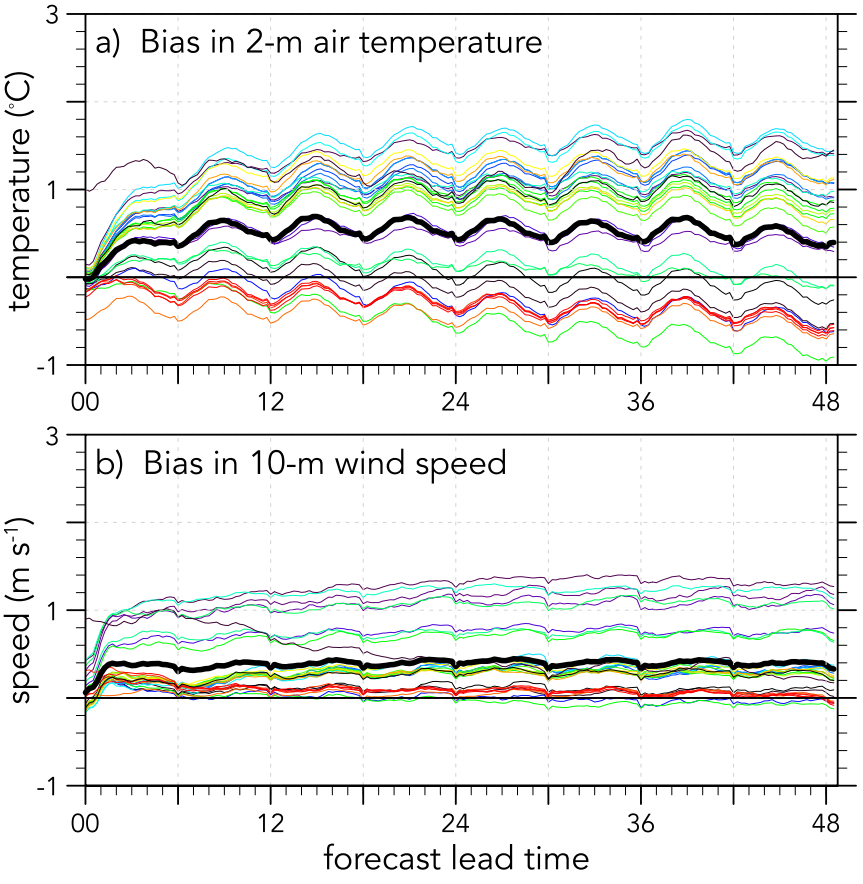

FIG. 9. Bias in (a) 2-m air temperature (°C) and (b) 10-m wind speed ($m\,s^{-1}$) as a function of forecast lead time on E-4DWX's domain 3 ($dx = 3.3$ km) in 2015 calculated from uncalibrated forecasts interpolated to 31 observation sites. The thick black line marks the ensemble mean. The thin colored lines mark individual ensemble members. Horizontal black lines mark zero bias.

prediction or any statistically small collection of predictions. The fact that the observation (red bar) in Fig. 6 *does not fall in the middle* of the bimodal PDF (solid black line) *does not indicate bias* any more than rolling a 2 or a 12 one time with a pair of six-sided dice indicates bias—that the dice are "loaded," in gamblers' parlance. Only through a statistically large number of predictions (or dice rolls) are biases revealed.

To get a better sense of how calibration can potentially alter a meteorologist's experience, we focus again on the sequence of 24-h forecasts in the gray inset of Fig. 5. Those forecasts valid for an unusual cool spell in June 2014 include several within a few degrees of 0°C. If a meteorologist were asked by a test director about the probability of temperatures approaching freezing, the differences between the uncalibrated and calibrated output from E-4DWX might be meaningful. Figure 7a presents E-4DWX's uncalibrated and calibrated 24-h forecasts of the probability of 2-m air temperature ≤0°–5°C in increments of 1°C. At the granularity of the ensemble, calibration has no effect on probabilistic forecasts of ≤0° and ≤1°, both of which fall in the lowest bin of 0%–10% chance (horizontal blue line and, hidden underneath it, red line in the first two of the six panels). There is an effect for ≤5°C (last of the five panels), but it is small. For ≤2°–4°C the effects are larger. At a threshold of 3°C (fourth of the six panels) calibration increases the probability from ≤10% to 25%. If we refer to the cost–loss model from section 1b, this difference equates to a calibrated loss threshold that is 40% of the uncalibrated loss threshold. If we imagine that a hypothetical test during the same cool spell is sensitive not just to temperature but also to wind, the effects of calibration are even more important (Fig. 7b). At a threshold of $5\,m\,s^{-1}$, calibration increases the forecast probability from ≤15% during the period to 65% at one point (middle panel in Fig. 7b).

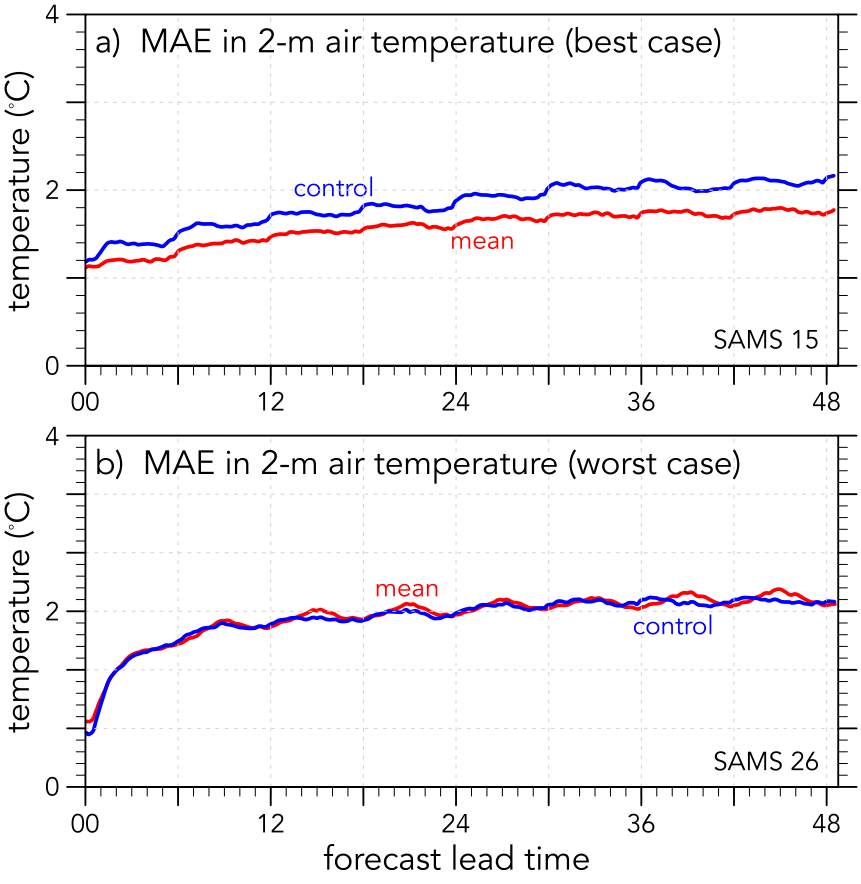

FIG. 10. MAE in 2-m air temperature (°C) at DPG's SAMS 15 and 26 from the E-4DWX ensemble mean (red) and the control member (blue) as a function of forecast lead time on E-4DWX's domain 3 (dx = 3.3 km) in 2015. Results from these two stations are representative of the sites where the mean provided (a) the *most* (labeled "*best* case") and (b) the *least* (labeled "*worst* case") improvement over the control member. Forecasts are uncalibrated.

b. *Overall skill of E-4DWX without calibration*

As mentioned earlier, at the time of writing, calibration is only applied to a small subset of E-4DWX's output: surface air pressure, temperature at 2 m, relative humidity at 2 m, and wind speed (including in terms of each horizontal component) at 10 m, and only at specific locations where observations are available. Many of the graphical products on which DPG's meteorologists rely are based on the ensemble's full 4D gridded output, so it is not sufficient that calibration produce highly skillful results. E-4DWX's predictions must be skillful even *without* calibration.

As expected, ensemble mean forecasts from E-4DWX have the most consistently low mean absolute error (MAE) over the entire 48 h of lead time (Fig. 8) compared with the forecasts of individual members, although for subsets of lead times there are a few members that produce slightly better forecasts of 2-m air temperature, but not of 10-m wind speed. (Recall from the introduction that the mean prediction by an ensemble tends to be more skillful than any individual member's prediction.) The mean forecast is biased high in temperature and wind speed (Fig. 9), but less so than the forecasts from many of the 30 members. The bias in the mean does not result from one or two extremely biased outlying members, but rather from a predominance of moderately biased members. Plots of RMSE and correlation (not shown) are consistent with plots of MAE: the ensemble mean performs best overall.

Before E-4DWX was deployed at DPG, meteorologists used only a deterministic version of 4DWX—which they still use in addition to the ensemble—so it is informative to compare the ensemble mean to the deterministic forecast from either of the two control members of the ensemble (Figs. 10–13). The configuration of the deterministic system includes a fourth domain with a

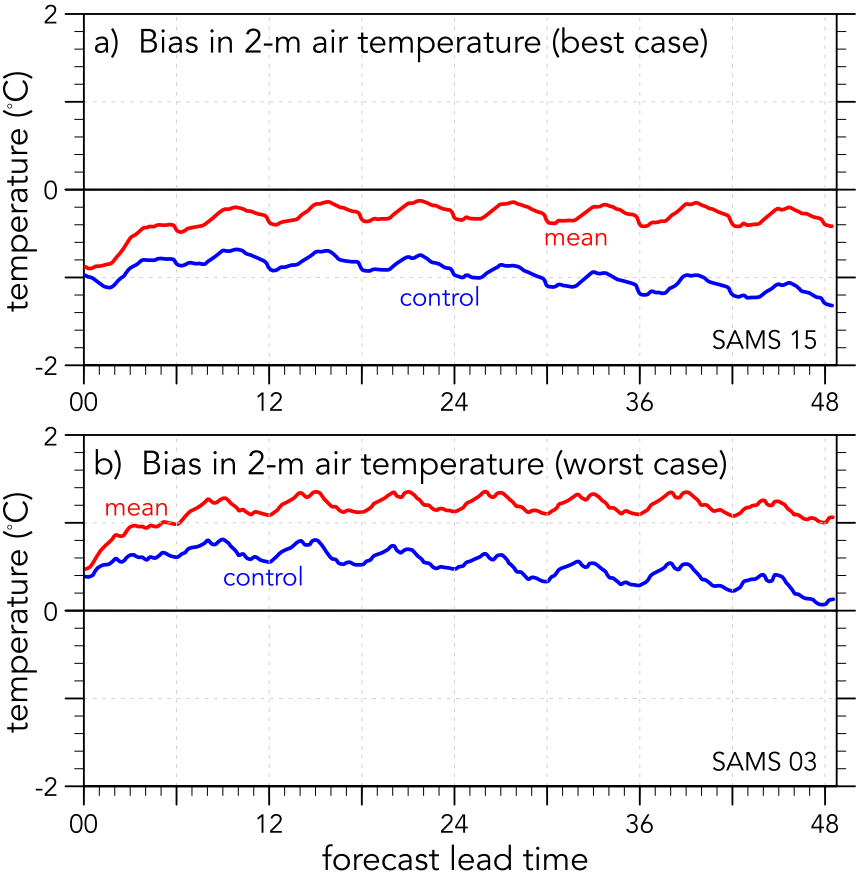

FIG. 11. As in Fig. 10, but bias in 2-m air temperature (°C) at DPG's SAMS 15 and 03. Horizontal black lines mark zero bias.

grid interval of 1.1 km, but its third domain is directly comparable to the finest domain of any of E-4DWX's members, including each control. Compared to the second of the controls (member 2 in Table 1), E-4DWX's ensemble mean has similar or lower MAEs in 2-m air temperature (Fig. 10) and 10-m wind speed (Fig. 12) at every SAMS station at DPG, but the mean is more biased than the control member at some stations (Figs. 11 and 13). The control member tends to predict lower 2-m air temperatures than does the ensemble mean, regardless of station, lead time, and time of day (not shown). It follows, then, that at stations where E-4DWX has a pronounced high bias, the control member has a bias closer to zero. For context on the control member's subtle, gradual cooling over the 48-h forecasts in Fig. 11, for which we do not yet have a complete explanation, please see the paper by Massey et al. (2016).

The undulations with a period of 6 h in some panels in Figs. 8 and 13 arise because at many stations E-4DWX's MAE and bias in 2-m air temperature and MAE in 10-m wind speed are functions of where the initialization time and the lead time fall during the 24 h of the day. E-4DWX's initialization times are fixed at 0000, 0600, 1200, and 1800 UTC. On these plots' axes (Figs. 8–13) each given lead time (e.g., 27 h from the four initializations on a given day) corresponds to four fixed points during the diurnal cycle (e.g., 0300, 0900, 1500, and 2100 UTC on the following day, respectively). As a result, the superposition of 1) varying analysis skill and forecast skill as a function of initialization time; 2) oscillating forecast skill as a function of where the valid time falls in the diurnal cycle, ignoring the slow shifts in the local solar cycle through the year; and 3) declining forecast skill as a function of lead time all combine to produce the 6-hourly undulations. A good illustration of the effect is Fig. 14, which shows the performance of the deterministic 4DWX at DPG in 2013. At a given lead time, the model's bias depends on the forecast's initialization time (Fig. 14a), and for a given forecast, the bias oscillates as a function of lead time, depending on where that lead time falls in the diurnal cycle (Fig. 14b). When all initializations are combined (heavy black line in

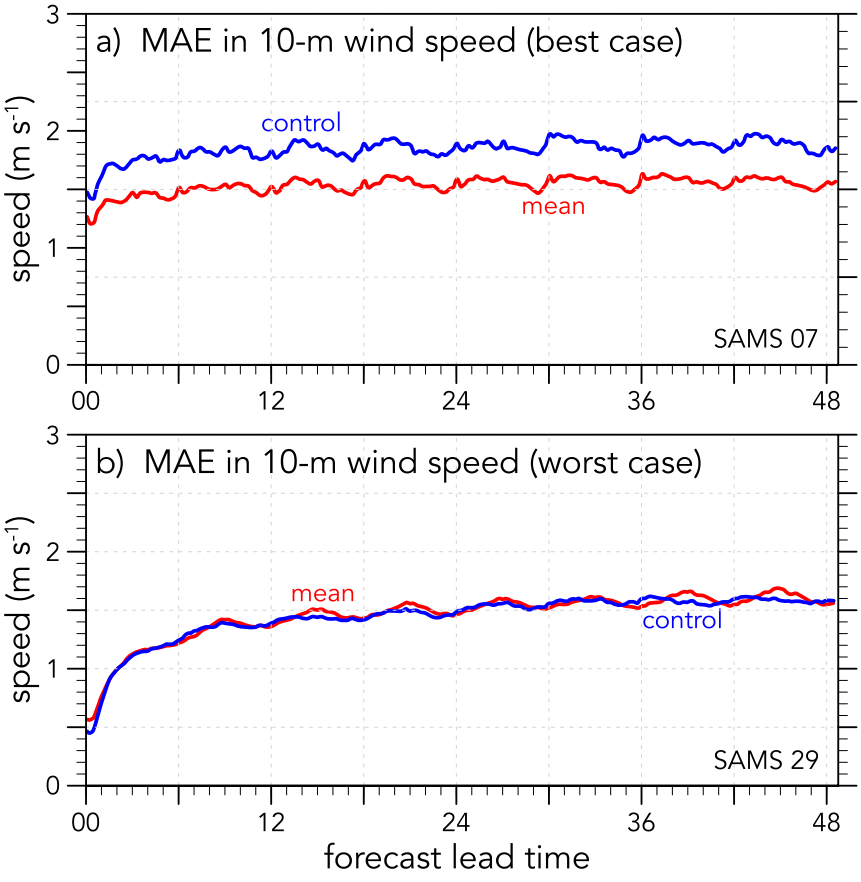

FIG. 12. As in Fig. 10, but MAE in 10-m wind speed ($m\,s^{-1}$) at DPG's SAMS 07 and 29.

Fig. 14a), the result is an undulating overall bias versus lead time.

Diurnally fluctuating biases in 2-m air temperature, as exemplified in Fig. 14b, are common in NWP forecasts at DPG and elsewhere in the semiarid and arid western United States (e.g., Cheng and Steenburgh 2005; Hart et al. 2005; Zheng et al. 2012). The deterministic 4DWX and the individual members of E-4DWX tend to underpredict the cool of the night (i.e., nighttime air temperature has a high bias) and the warmth of the day (i.e., daytime air temperature has a low bias). Sources of this bias include how the thermal conductivity of soil is characterized in the model code (Massey et al. 2014) and how moisture in the upper levels of soil is overestimated in land surface analyses (Massey et al. 2016).

4. Products

Each cycle of E-4DWX generates nearly half a terabyte of output. How the vast amount of data from E-4DWX or any other EPS is organized and provided to meteorologists determines the system's success or failure as much as does any of an EPS's other components (Buizza et al. 2005). To generate useful products from E-4DWX, during each forecast cycle an elaborate set of algorithms is run on the computational nodes of the host Linux cluster. By the time the combinations of ensemble members, domains, lead times, atmospheric variables, vertical levels, etc. are considered, the output products number in the hundreds of thousands per cycle. The products' development has been guided by the meteorologists at DPG and by the testing that they support because, as Palmer (2002) recommended, EPS predictions should be expressed in terms of the most relevant user variables. Graphics (weather maps, time series, etc.) are generated with the NCAR Command Language (NCL) and with Read/Interpolate/Plot (RIP). These graphics are then transferred to what is known as a 4DWX Data Application Server (DAS), a Linux node that is separate from the cluster that hosts the NWP components of E-4DWX. The DAS includes a web server that allows registered users to view products from E-4DWX via a collection of web pages.

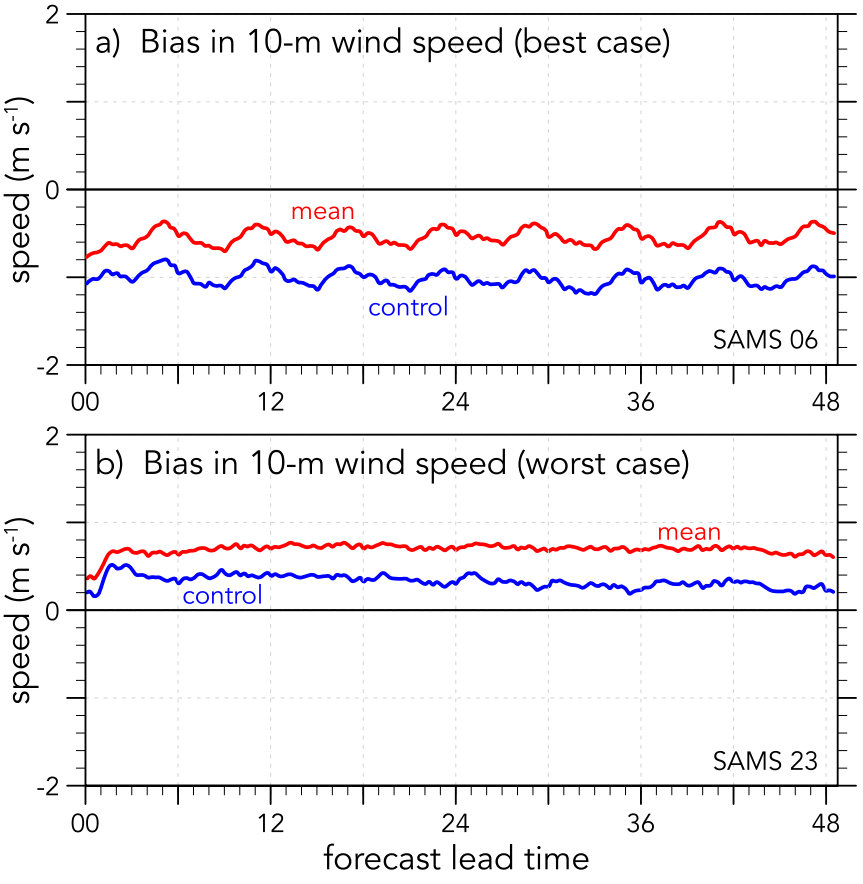

FIG. 13. As in Fig. 10, but bias in 10-m wind speed ($m\,s^{-1}$) at DPG's SAMS 06 and 23. Horizontal black lines mark zero bias.

a. Weather maps

1) ENSEMBLE MEAN

The most frequently used products from E-4DWX are based on the ensemble mean. The mean is available to meteorologists through weather maps that present the same fields found on maps derived from the control members (not shown). Therefore, the variables and display formats on these maps are familiar to the meteorologists, but the specific guidance itself is more skillful than if based only on a single deterministic run (see sections 2 and 3).

Standard surface and upper-level maps are generated by E-4DWX. In addition, profiles of wind and temperature for common test locations are included with the ensemble-mean suite of products. Conditions in the lower troposphere are what most affect testing at DPG, so products include a series of plots displaying wind speed and direction from altitudes of 10 m to 2 km (not shown).

2) EXCEEDANCE

As emphasized elsewhere in this paper, many of the tests at DPG are sensitive to weather in a binary fashion. Conditions beyond certain thresholds make tests impossible, invalid, dangerous, or otherwise unacceptable. In such cases guidance from E-4DWX is most valuable when cast as likelihoods that thresholds will be exceeded. Probabilities are generated by counting the number of ensemble members that meet or exceed a threshold, then dividing the count by the total number of ensemble members. Figure 15a is an example. Probabilistic output from exceedance plots also allows meteorologists to infer how uncertainty in E-4DWX predictions varies in space and time.

3) MEAN AND STANDARD DEVIATION

Variability in the predictions among ensemble members can be conveyed through plots of standard deviation from the mean (Fig. 15b). Higher standard deviations indicate areas of more uncertainty where model forecasts tend to be less reliable.

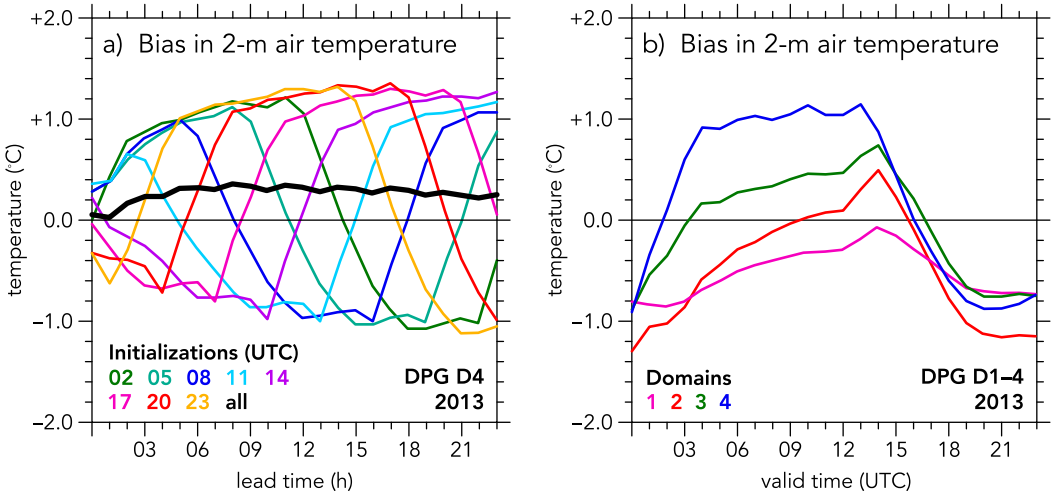

FIG. 14. Deterministic 4DWX's bias in 2-m air temperature (°C) in 2013 (a) within domain 4 (dx = 1.1 km) as a function of lead time (h) from eight separate initialization times (colors; see key) and (b) within domains 1 (dx = 30 km), 2 (dx = 10 km), 3 (dx = 3.3 km), and 4 (dx = 1.1 km) as a function of valid time (UTC) from all initialization times combined. Biases are calculated from uncalibrated forecasts initialized every 3 h between 0200 and 2300 UTC each day, interpolated to all observation sites.

4) SPAGHETTI PLOTS

Output from all ensemble members can be plotted spatially on the whimsically named *spaghetti* plot (Fig. 16a). These plots, too, provide guidance on uncertainty. When lines are tightly spaced around similar solutions from a collection of ensemble members, confidence in predictions is relatively high. Spaghetti plots are especially useful for depicting when the ensemble's overall spread comprises several spatial clusters of distinct solutions.

b. Meteograms

Most tests at DPG occur at specific locations where weather stations have been installed. For these sites, E-4DWX provides output from all ensemble members as *meteorograms* (commonly shortened in the community to *meteograms*). An example is Fig. 16b, in which each ensemble quantile's calibrated prediction of 10-m wind speed is plotted as a function of time. The ensemble mean (green) and spread of ±1 standard deviation (yellow) are then superposed on the individual forecasts. Other meteograms present 2-m temperature, 2-m relative humidity, altitude of the ABL, 10-m wind roses, and 1-h accumulated precipitation (not shown). E-4DWX also provides data in other formats very similar to a meteogram, such as time series of box-and-whisker plots (not shown). Recent observations and predictions are often presented together on meteograms (black line in Fig. 16b) so meteorologists can consider E-4DWX's recent performance when developing their weather guidance.

c. Wind roses

One means of conveying E-4DWX's distribution of wind predictions is through a wind rose (Fig. 17). On this type of plot, the azimuth corresponds to the direction of origin, discretized into 16 bins or sectors (e.g., a northerly wind is plotted in the sector centered on azimuth 0°). Within each sector, the radial dimension of shading from nearer the center of the rose outward indicates the percentage of members predicting wind direction within that sector. Range rings mark every 20% of the ensemble's membership. Wind speed predictions are displayed in the sectors according to the hues on a color bar. In Fig. 17 for example, the ensemble predicts easterly winds from 0 to 6 m s^{-1} from the initialization time of 0600 UTC until 1200 UTC, after which the winds back to northwesterly. Winds weaken between 0000 and 0300 UTC on day 2, then turn southerly and increase to 4–10 m s^{-1} among the members predicting the highest speeds.

5. Summary

This paper is an overview of E-4DWX, an ensemble prediction system (EPS) that has been used for test support at Dugway Proving Ground (DPG) in northwestern Utah since 2007. In 2014, three more U.S. Army test facilities adopted a slightly lower-resolution version of the system: White Sands Missile Range, Yuma Proving Ground, and Electronic Proving Ground.

The 30 operational members of E-4DWX each produce forecasts of 48 h every 6 h on a 272-processor

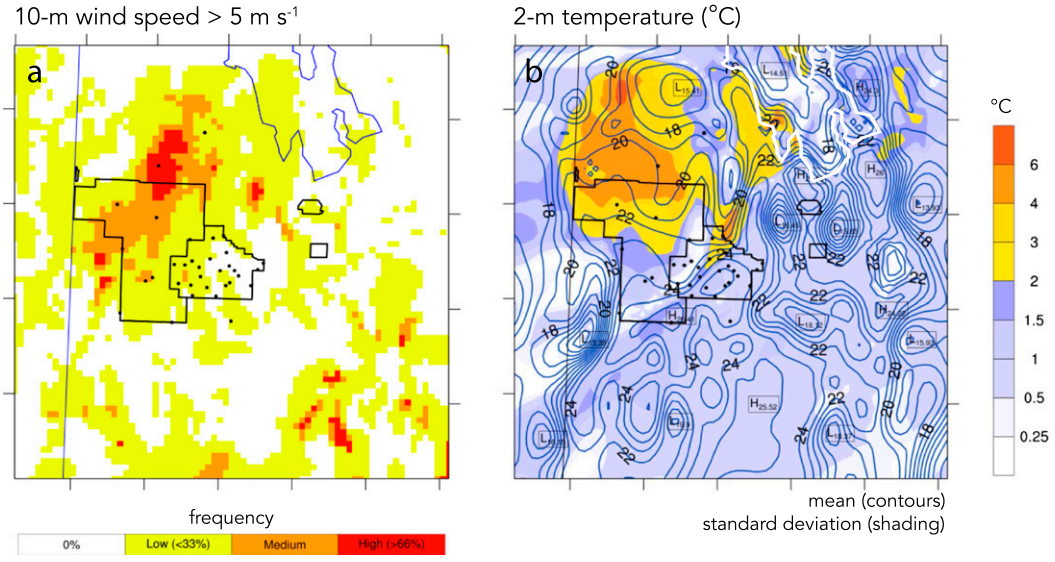

FIG. 15. Instantaneous forecasts of (a) the uncalibrated probability of 10-m wind speed exceeding 5 m s^{-1} and (b) the uncalibrated mean (blue contours with black labels) and standard deviations (colored shading) of 2-m air temperature (°C). In (a) the colors mark grid cells in which 0 (white), 1–9 (yellow), 10–20 (orange), and more than 21 (red) members of the 30-member E-4DWX predicted that the wind speed threshold would be exceeded at the valid time. In (b) the local maxima (H) and minima (L) are indicated. Both forecasts are on domain 3 (dx = 3.3 km) at unspecified times and dates. DPG is outlined in black, the Great Salt Lake is blue in (a) and white in (b), and black dots mark sites of observations at 2 and 10 m.

high performance computer at DPG. All members are based on the WRF Model. Spread is generated through variations in members' initial-, lateral-, and lower-boundary conditions, as well as in physical parameterizations.

E-4DWX's calibration is innovative. For locations where observations are available, the most useful near-surface variables are dynamically calibrated through a combination of logistic and quantile regressions, explicitly conditioned on ensemble dispersion. Calibration

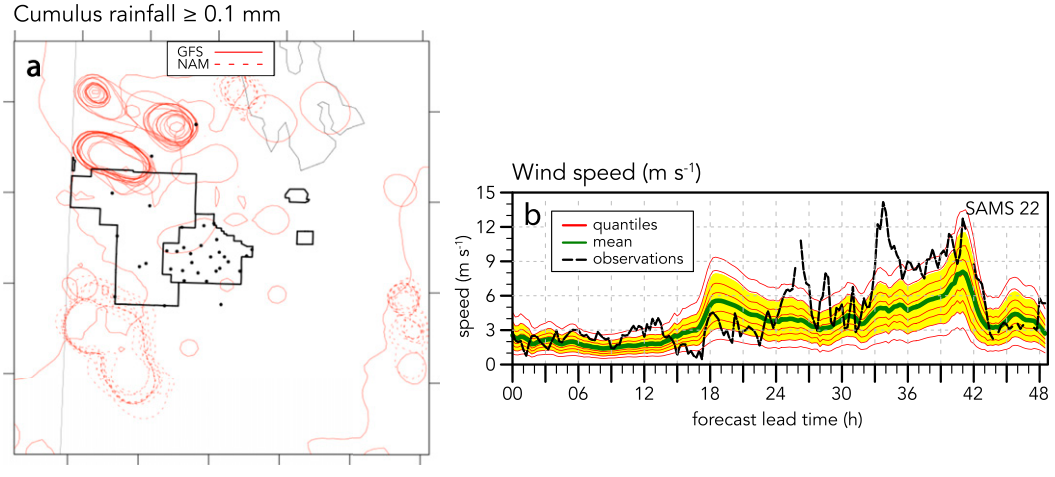

FIG. 16. Forecasts of (a) regions of uncalibrated precipitation ≥ 0.1 mm (liquid equivalent) from the cumulus scheme in the previous hour (red contours) by individual ensemble members whose initial and boundary conditions were from the NAM (dashed) or GFS (solid), and (b) calibrated 10-m wind speed (m s^{-1}) as a function of forecast lead time at SAMS 22 at DPG. In (a) DPG is outlined in thick black, the Great Salt Lake is in thin black, and black dots mark sites of observations at 2 and 10 m. In (b), the E-4DWX quantiles are in red, the ensemble mean is in green, ±1 standard deviation is in yellow, and observations are in black. Forecasts are on domain 3 (dx = 3.3 km) at unspecified times and dates.

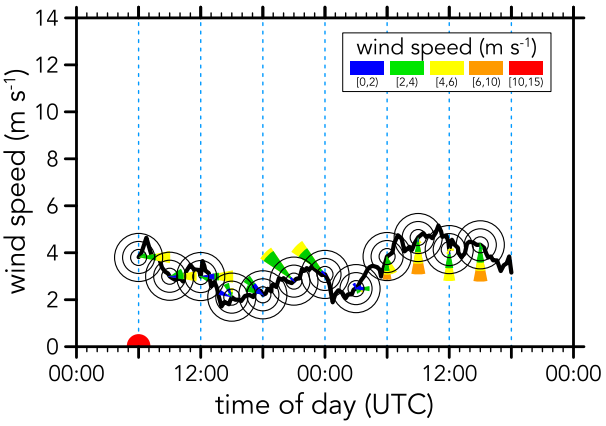

FIG. 17. Time series of E-4DWX's mean 10-m wind speed (thick black line; $m\,s^{-1}$) and ensemble members' wind directions (wind roses) at an unspecified site. This 48-h uncalibrated forecast (plus 6 h of data assimilation at the start of the period, to the left of the red semicircle along the x axis) is on domain 3 (dx = 3.3 km) on an unspecified day during the spring. Each wind rose shows for a given time of day (UTC) the percentage of ensemble members (range rings at intervals of 20%) that predict ranges of wind speed (colors) from specific directions (sectors 22.5° wide). Each rose's position along the y axis marks the ensemble's mean wind speed at that time.

is applied within a cross-validation framework to minimize the likelihood of overfitting. For the calibrated variables, E-4DWX is realistically dispersive and unbiased. Rank histograms following calibration against a reference dataset generally fall within the 95% confidence bounds for statistically perfect calibration. For individual ensemble forecasts, the most noticeable effects of calibration are to broaden the spread among members and to shift the mean of their distribution toward a less biased value.

Without calibration, E-4DWX's ensemble means of 2-m temperature and 10-m wind speed (among the variables of most concern to meteorologists at DPG) are skillful but tend to be biased high, owing to a predominance of moderately biased members, not a few extremely biased members. During the full 48 h of lead times, the ensemble mean consistently has lower MAEs overall than even E-4DWX's best-performing members. For subsets of lead times, there are a few members that sometimes produce slightly better forecasts of 2-m air temperature, but not of 10-m wind speed. The ensemble mean also has consistently lower RMSEs and higher correlations than individual members.

Each cycle of E-4DWX generates nearly half a terabyte of output that is translated into graphical products whose design is based on requests from meteorologists. Products display means, standard deviations, or fractions of the ensemble exceeding a threshold, as well as predictions from individual members of the ensemble.

Although likelihood and uncertainty from an EPS are valuable, many test participants are inexperienced with weather guidance in those terms and still prefer binary go/no-go (yes/no) recommendations from a meteorologist. Meteorologists at DPG are working to determine how best to use and communicate the confidence and uncertainty information from E-4DWX. In our view, calibrated probabilistic guidance and some notion of the repercussions of action or inaction are requisite for using weather forecasts to make rational, objective decisions.

Acknowledgments. This work was funded by the U.S. Army Test and Evaluation Command through an Interagency Agreement with the National Science Foundation, which sponsors the National Center for Atmospheric Research. Thanks to the late Thomas Warner for treasured mentorship, James Bowers and Danielle Terrin for their support during the project's earlier stages, and to Jessica Knight for her support now. DPG's funding for the HPC that hosts E-4DWX has been fundamental to the success of the project. Choices in the design of E-4DWX are based on valuable input from the following current and former meteorologists at DPG: William Ahue, Edward Argenta, Jeffrey Braun, Jarrett Claiborne, Trisha Gabbert, Matthew Jeglum, Carissa Klemmer, Susan Krippner, Matthew Lloyd, and Daniel Ruth. Constructive comments from three reviewers helped us improve the manuscript following its initial submission. Most figures were initially generated with the NCAR Command Language (NCL), Read/Interpolate/Plot (RIP), Interactive Data Language (IDL), and MATLAB. Most figures were further refined with Adobe Illustrator (AI). Thanks to Donny Storwald for providing the core of Fig. 2 and to Joe Grim for generating Fig. 14. Code for calculating LR and QR were based on code from The R Project (http://www.r-project.org).

REFERENCES

Atger, F., 1999: The skill of ensemble prediction systems. *Mon. Wea. Rev.*, **127**, 1941–1953, https://doi.org/10.1175/1520-0493(1999)127<1941:TSOEPS>2.0.CO;2.

——, 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.*, **131**, 1509–1523, https://doi.org/10.1175//1520-0493(2003)131<1509:SAIVOT>2.0.CO;2.

Bentzien, S., and P. Friederichs, 2012: Generating and calibrating probabilistic quantitative precipitation forecasts from the high-resolution NWP model COSMO-DE. *Wea. Forecasting*, **27**, 988–1002, https://doi.org/10.1175/WAF-D-11-00101.1.

Berner, J., K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder, 2015: Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Wea. Rev.*, **143**, 1295–1320, https://doi.org/10.1175/MWR-D-14-00091.1.

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.*, **135**, 1386–1402, https://doi.org/10.1175/MWR3341.1.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Mon. Wea. Rev.*, **140**, 3706–3721, https://doi.org/10.1175/MWR-D-12-00031.1.

Bremnes, J. B., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, https://doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Buizza, R., 2008: The value of probabilistic prediction. *Atmos. Sci. Lett.*, **9**, 36–42, https://doi.org/10.1002/asl.170.

——, P. L. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.*, **133**, 1076–1097, https://doi.org/10.1175/MWR2905.1.

Cheng, W. Y., and W. J. Steenburgh, 2005: Evaluation of surface sensible weather forecasts by the WRF and the Eta models over the western United States. *Wea. Forecasting*, **20**, 812–821, https://doi.org/10.1175/WAF885.1.

Clark, A. J., W. A. Gallus Jr., M. Xue, and F. Kong, 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Wea. Forecasting*, **24**, 1121–1140, https://doi.org/10.1175/2009WAF2222222.1.

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, https://doi.org/10.1175/MWR-D-12-00281.1.

Diomede, T., C. Marsigli, A. Montani, F. Nerozzi, and T. Paccagnella, 2014: Calibration of limited-area ensemble precipitation forecasts for hydrological predictions. *Mon. Wea. Rev.*, **142**, 2176–2197, https://doi.org/10.1175/MWR-D-13-00071.1.

Du, J., S. L. Mullen, and F. Sanders, 1997: Short-range ensemble forecasting of quantitative precipitation. *Mon. Wea. Rev.*, **125**, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2.

Ebert, E. E., 2001: Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Wea. Rev.*, **129**, 2461–2480, https://doi.org/10.1175/1520-0493(2001)129<2461:AOAPMS>2.0.CO;2.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting*, **13**, 1132–1147, https://doi.org/10.1175/1520-0434(1998)013<1132:CPQPFB>2.0.CO;2.

——, and C. F. Mass, 2005: Aspects of effective mesoscale, short-range ensemble forecasting. *Wea. Forecasting*, **20**, 328–350, https://doi.org/10.1175/WAF843.1.

——, H. R. Glahn, T. M. Hamill, S. Joslyn, W. M. Lapenta, and C. F. Mass, 2010: National mesoscale probabilistic prediction: Status and the way forward. White Paper Rep. from the National Workshop on Mesoscale Probabilistic Prediction, NWS, 24 pp., http://www.nws.noaa.gov/ost/NMPP_white_paper_28May10_with%20sig%20page.pdf.

Fernando, H. J. S., and Coauthors, 2015: The MATERHORN: Unraveling the intricacies of mountain weather. *Bull. Amer. Meteor. Soc.*, **96**, 1945–1967, https://doi.org/10.1175/BAMS-D-13-00131.1.

Fortin, V., A.-C. Favre, and M. Saïd, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.*, **132**, 1349–1369, https://doi.org/10.1256/qj.05.167.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268, https://doi.org/10.1175/2008MWR2569.1.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Grachev, A. A., L. S. Leo, S. D. Sabatino, H. J. S. Fernando, E. R. Pardyjak, and C. W. Fairall, 2016: Structure of turbulence in katabatic flows below and above the wind-speed maximum. *Bound.-Layer Meteor.*, **159**, 469–494, https://doi.org/10.1007/s10546-015-0034-8.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, https://doi.org/10.1175/2007MWR2410.1.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, https://doi.org/10.1002/qj.1895.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, https://doi.org/10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

——, S. L. Mullen, C. Snyder, D. P. Baumhefner, and Z. Toth, 2000: Ensemble forecasting in the short to medium range: Report from a workshop. *Bull. Amer. Meteor. Soc.*, **81**, 2653–2664, https://doi.org/10.1175/1520-0477(2000)081<2653:EFITST>2.3.CO;2.

——, J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, https://doi.org/10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, https://doi.org/10.1175/2007MWR2411.1.

Hart, K. A., W. J. Steenburgh, and D. J. Onton, 2005: Model forecast improvements with decreased horizontal grid spacing over finescale intermountain orography during the 2002 Olympic Winter Games. *Wea. Forecasting*, **20**, 558–576, https://doi.org/10.1175/WAF865.1.

Hilliker, J. L., and J. M. Fritsch, 1999: An observations-based statistical system for warm-season hourly probabilistic forecasts of low ceiling at the San Francisco International Airport. *J. Appl. Meteor.*, **38**, 1692–1705, https://doi.org/10.1175/1520-0450(1999)038<1692:AOBSSF>2.0.CO;2.

Hohenegger, C., and C. Schär, 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Amer. Meteor. Soc.*, **88**, 1783–1793, https://doi.org/10.1175/BAMS-88-11-1783.

Hopson, T. M., 2014: Assessing the ensemble spread–error relationship. *Mon. Wea. Rev.*, **142**, 1125–1142, https://doi.org/10.1175/MWR-D-12-00111.1.

——, and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods

of 2003–07. *J. Hydrometeor.*, **11**, 618–641, https://doi.org/10.1175/2009JHM1006.1.

Iyer, E. R., A. J. Clark, M. Xue, and F. Kong, 2016: A comparison of 36–60-h precipitation forecasts from convection-allowing and convection-parameterizing ensembles. *Wea. Forecasting*, **31**, 647–661, https://doi.org/10.1175/WAF-D-15-0143.1.

Jeglum, M. E., and S. W. Hoch, 2016: Multiscale characteristics of surface winds in an area of complex terrain in northwest Utah. *J. Appl. Meteor. Climatol.*, **55**, 1549–1563, https://doi.org/10.1175/JAMC-D-15-0313.1.

——, ——, D. D. Jensen, R. Dimitrova, and Z. Silver, 2017: Large temperature fluctuations due to cold-air pool displacement along the lee slope of a desert mountain. *J. Appl. Meteor. Climatol.*, **56**, 1083–1098, https://doi.org/10.1175/JAMC-D-16-0202.1.

Johnson, A., and X. Wang, 2012: Verification and calibration of neighborhood and object-based probabilistic precipitation forecasts from a multimodel convection-allowing ensemble. *Mon. Wea. Rev.*, **140**, 3054–3077, https://doi.org/10.1175/MWR-D-11-00356.1.

——, and Coauthors, 2014: Multiscale characteristics and evolution of perturbations for warm season convection-allowing precipitation forecasts: Dependence on background flow and method of perturbation. *Mon. Wea. Rev.*, **142**, 1053–1073, https://doi.org/10.1175/MWR-D-13-00204.1.

Junk, C., L. Delle Monache, and S. Alessandrini, 2015: Analog-based ensemble model output statistics. *Mon. Wea. Rev.*, **143**, 2909–2917, https://doi.org/10.1175/MWR-D-15-0095.1.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability.* 1st ed. Cambridge University Press, 364 pp.

Kleiber, W., A. E. Raftery, J. Baars, T. Gneiting, C. F. Mass, and E. Grimit, 2011: Locally calibrated probabilistic temperature forecasting using geostatistical model averaging and local Bayesian model averaging. *Mon. Wea. Rev.*, **139**, 2630–2649, https://doi.org/10.1175/2010MWR3511.1.

Koenker, R., and G. Bassett, 1978: Regression quantiles. *Econometrica*, **46**, 33–50, https://doi.org/10.2307/1913643.

Krzysztofowicz, R., and A. A. Sigrest, 1999: Comparative verification of guidance and local quantitative precipitation forecasts: Calibration analyses. *Wea. Forecasting*, **14**, 443–454, https://doi.org/10.1175/1520-0434(1999)014<0443:CVOGAL>2.0.CO;2.

Lehner, M., and Coauthors, 2015: A case study of the nocturnal boundary layer evolution on a slope at the foot of a desert mountain. *J. Appl. Meteor. Climatol.*, **54**, 732–751, https://doi.org/10.1175/JAMC-D-14-0223.1.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409–418, https://doi.org/10.1175/1520-0493(1974)102<0409:TSOMCF>2.0.CO;2.

Liu, Y., J. P. Hacker, T. T. Warner, and S. Swerdlin, 2007: A WRF- and MM5-based 4-D mesoscale ensemble data analysis and prediction system (E-RTFDDA) developed for ATEC operational applications. *18th Conf. on Numerical Weather Prediction*, Park City, UT, Amer. Meteor. Soc., 7B.7, https://ams.confex.com/ams/pdfpapers/124739.pdf.

——, and Coauthors, 2008: The operational mesogamma-scale analysis and forecast system of the U.S. Army Test and Evaluation Command. Part I: Overview of the modeling system, the forecast products, and how the products are used. *J. Appl. Meteor. Climatol.*, **47**, 1077–1092, https://doi.org/10.1175/2007JAMC1653.1.

Ma, J., Y. Zhu, R. Wobus, and P. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.*, **29**, 782–794, https://doi.org/10.1007/s00376-012-1249-y.

Massey, J. D., W. J. Steenburgh, S. W. Hoch, and J. C. Knievel, 2014: Sensitivity of near-surface temperature forecasts to soil properties over a sparsely vegetated dryland region. *J. Appl. Meteor. Climatol.*, **53**, 1976–1995, https://doi.org/10.1175/JAMC-D-13-0362.1.

——, ——, J. C. Knievel, and W. Y. Y. Cheng, 2016: Regional soil moisture biases and their influence on WRF Model temperature forecasts over the Intermountain West. *Wea. Forecasting*, **31**, 197–216, https://doi.org/10.1175/WAF-D-15-0073.1.

——, ——, S. W. Hoch, and D. D. Jensen, 2017: Simulated and observed surface energy fluxes and resulting playa breezes during the MATERHORN field campaigns. *J. Appl. Meteor. Climatol.*, **56**, 915–935, https://doi.org/10.1175/JAMC-D-16-0161.1.

McDonald, A., 1997: Lateral boundary conditions for operational regional forecast models; a review. HIRLAM Tech. Rep. 32, 31 pp., http://hirlam.org/index.php/hirlam-documentation/doc_view/1323-hirlam-technical-report-no-32.

Messner, J. W., G. J. Mayr, D. S. Wilks, and A. Zeileis, 2014a: Extending extended logistic regression: Extended versus separate versus ordered versus censored. *Mon. Wea. Rev.*, **142**, 3003–3014, https://doi.org/10.1175/MWR-D-13-00355.1.

——, ——, A. Zeileis, and D. S. Wilks, 2014b: Heteroscedastic extended logistic regression for postprocessing of ensemble guidance. *Mon. Wea. Rev.*, **142**, 448–456, https://doi.org/10.1175/MWR-D-13-00271.1.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF Ensemble Prediction System: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Murphy, A. H., 1977: The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. *Mon. Wea. Rev.*, **105**, 803–816, https://doi.org/10.1175/1520-0493(1977)105<0803:TVOCCA>2.0.CO;2.

Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.*, **9**, 307–315, https://doi.org/10.1017/S1350482702003043.

Nehrkorn, T., B. Woods, T. Auligné, and R. N. Hoffman, 2014: Application of feature calibration and alignment to high-resolution analysis: Examples using observations sensitive to cloud and water vapor. *Mon. Wea. Rev.*, **142**, 686–702, https://doi.org/10.1175/MWR-D-13-00164.1.

Nutter, P., D. Stensrud, and M. Xue, 2004: Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Wea. Rev.*, **132**, 2358–2377, https://doi.org/10.1175/1520-0493(2004)132<2358:EOCRAT>2.0.CO;2.

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.

Pielke, R. A., Sr., 2002: *Mesoscale Meteorological Modeling.* 2nd ed. Academic Press, 676 pp.

Portnoy, S., and R. Koenker, 1997: The Gaussian hare and the Laplacian tortoise: Computability of squared-error versus absolute-error estimators. *Stat. Sci.*, **12**, 279–300, https://doi.org/10.1214/ss/1030037960.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rife, D. L., T. T. Warner, F. Chen, and E. G. Astling, 2002: Mechanisms for diurnal boundary layer circulations in the Great Basin Desert. *Mon. Wea. Rev.*, **130**, 921–938, https://doi.org/10.1175/1520-0493(2002)130<0921:MFDBLC>2.0.CO;2.

Romine, G. S., C. S. Schwartz, J. Berner, K. R. Fossell, C. Snyder, J. L. Anderson, and M. L. Weisman, 2014: Representing forecast error in a convection-permitting ensemble system. *Mon. Wea. Rev.*, **142**, 4519–4541, https://doi.org/10.1175/MWR-D-14-00100.1.

Roulin, E., and S. Vannitsem, 2012: Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Wea. Rev.*, **140**, 874–888, https://doi.org/10.1175/MWR-D-11-00062.1.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, https://doi.org/10.3402/tellusa.v55i1.12082.

Schwartz, C. S., G. S. Romine, K. R. Smith, and M. L. Weisman, 2014: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter. *Wea. Forecasting*, **29**, 1295–1318, https://doi.org/10.1175/WAF-D-13-00145.1.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp., http://dx.doi.org/10.5065/D68S4MVH.

Sloughter, J. M. L., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, https://doi.org/10.1175/MWR3441.1.

Stauffer, D. R., and N. L. Seaman, 1994: Multiscale four-dimensional data assimilation. *J. Appl. Meteor.*, **33**, 416–434, https://doi.org/10.1175/1520-0450(1994)033<0416:MFDDA>2.0.CO;2.

Tokdar, S. T., and J. B. Kadane, 2012: Simultaneous linear quantile regression: A semiparametric Bayesian approach. *Bayesian Anal.*, **7**, 51–72, https://doi.org/10.1214/12-BA702.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.*, **74**, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2.

——, and ——, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting*, **16**, 463–477, https://doi.org/10.1175/1520-0434(2001)016<0463:TUOETI>2.0.CO;2.

Unger, D. A., H. Van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379, https://doi.org/10.1175/2008MWR2605.1.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, https://doi.org/10.1256/qj.04.120.

Warner, T. T., 2011: *Numerical Weather and Climate Prediction.* Cambridge University Press, 526 pp.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl.*, **13**, 243–256, https://doi.org/10.1017/S1350482706002192.

——, 2006b: *Statistical Methods in the Atmopsheric Sciences.* 2nd ed. 467 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/10.1175/MWR3402.1.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian Ensemble Prediction System using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 1364–1385, https://doi.org/10.1175/MWR3347.1.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H.-M. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303, https://doi.org/10.1175/2007WAF2006114.1.

Zheng, W., H. Wei, Z. Wang, X. Zeng, J. Meng, M. Ek, K. Mitchell, and J. Derber, 2012: Improvement of daytime land surface skin temperature over arid regions in the NCEP GFS model and its impact on satellite data assimilation: LST improvement and data assimilation. *J. Geophys. Res.*, **117**, D06117, doi:10.1029/2011JD015901.