Improving Ensemble Weather Prediction System Initialization: Disentangling the Contributions from Model Systematic Errors and Initial Perturbation Size

THOMAS M. HAMILL[a] AND MICHAEL SCHEUERER[a,b]

[a] *NOAA/Earth System Research Laboratory, Physical Sciences Laboratory, Boulder, Colorado*
[b] *Cooperative Institute for Research in the Environmental Sciences, University of Colorado Boulder, Boulder, Colorado*

(Manuscript received 20 April 2020, in final form 7 October 2020)

ABSTRACT: Characteristics of the European Centre for Medium-Range Weather Forecast's (ECMWF's) 0000 UTC diagnosed 2-m temperatures (T_{2m}) from 4D-Var and global ensemble forecasts initial conditions were examined in 2018 over the contiguous United States at 1/2° grid spacing. These were compared against independently generated, upscaled high-resolution T_{2m} analyses that were created with a somewhat novel data assimilation methodology, an extension of classical optimal interpolation (OI) to surface data analysis. The analysis used a high-resolution, spatially detailed climatological background and was statistically unbiased. Differences of the ECMWF 4D-Var T_{2m} initial states from the upscaled OI reference were decomposed into a systematic component and a residual component. The systematic component was determined by applying a temporal smoothing to the time series of differences between the ECMWF T_{2m} analyses and the OI analyses. Systematic errors at 0000 UTC were commonly 1 K or more and larger in the mountainous western United States, with the ECMWF analyses cooler than the reference. The residual error is regarded as random in character and should be statistically consistent with the spread of the ensemble of initial conditions after inclusion of OI analysis uncertainty. This analysis uncertainty was large in the western United States, complicating interpretation. There were some areas suggestive of an overspread initial ensemble, with others underspread. Assimilation of more observations in the reference OI analysis would reduce analysis uncertainty, facilitating more conclusive determination of initial-condition ensemble spread characteristics.

KEYWORDS: Error analysis; Numerical analysis/modeling; Statistical techniques; Data assimilation; Ensembles

1. Introduction

Ensembles of numerical weather forecasts are commonly synthesized to provide users with state-dependent estimates of the weather forecast uncertainty (Palmer 2006; Warner 2011; Buizza 2018). These facilitate improved decisions relative to making them based on deterministic forecast guidance (Richardson 2000; Zhu et al. 2002). Great strides have been made in generating more skillful and reliable ensemble predictions in recent years; for example, see Buizza (2019) and Palmer (2019a). Still, sensible weather elements such as precipitation and 2-m temperature (T_{2m}) over land are still very challenging to simulate with skill and reliability in deterministic and ensemble forecasts (Sutton et al. 2006; Hamill and Whitaker 2007; Lavaysse et al. 2013; Tennant and Beare 2014; Gehne et al. 2019).

Suppose a developer has generated common diagnostics such as rank histograms on T_{2m}. These exhibit a U shape, which are typically interpreted as indicating a lack of spread in the ensemble. As discussed in Hamill (2001), there are other possible interpretations that include combinations of ensembles with warm biases in some samples and cold biases in others. The rank histogram alone does not provide enough information to indicate which aspects of an ensemble prediction system should be improved, and we consider other diagnostics to help determine whether deficiencies were related to systematic errors in the underlying deterministic predictions or deficiencies in the ensemble construction.

Ideally, for an ensemble prediction system we desire the individual members' predictions to be unbiased, with no systematic differences from the truth. This is meant both in a univariate sense, such as no bias in the verification of a member against an unbiased point observation, and in a multivariate sense, such as providing unbiased forecasts of the propagation speed and amplitudes of low-frequency phenomena such as the Madden–Julian oscillation (Zhang 2013 and references therein). To approach this goal, prediction centers like ECMWF use numerical weather prediction systems with advanced data assimilation, numerical methods, physical parameterizations, and land–atmosphere–ocean–sea ice couplings (e.g., Bonavita et al. 2016; Juricke et al. 2018; Ahlgrimm et al. 2018; Beljaars et al. 2018; Haiden et al. 2018a). Additionally, its ensemble prediction system should also forecast the state-dependent uncertainty as accurately as possible, with maximal sharpness subject to calibration (Gneiting et al. 2007). For this, the initial ensemble-state estimates should represent draws from the distribution of plausible analysis states (e.g., Houtekamer and Mitchell 1998, 2001, 2005; Buehner et al. 2010a,b; Bonavita et al. 2011). The ensemble prediction system must also realistically simulate the growth of random errors during the forecast due to

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/MWR-D-20-0119.s1.

Corresponding author: Thomas M. Hamill, Tom.Hamill@noaa.gov

DOI: 10.1175/MWR-D-20-0119.1

chaos (Lorenz 1963; Molteni et al. 1996; Palmer 2006) and the growth of forecast uncertainty due to model simplifications (Buizza et al. 1999).

A challenge with developing model diagnostics to isolate error sources is that modifications to address deterministic systematic error can affect ensemble spread, and vice versa. For example, the introduction of stochastic physics to address the model uncertainty may also improve the prediction system's systematic errors (Weisheimer et al. 2014; Christensen et al. 2017; Palmer 2001, 2019b). Further, an improved forecast model used in the ensemble will result in a better fit of the background to the observations, reducing the spread of the estimated initial conditions and the spread in the subsequent short- to medium-range ensembles. Numerical weather prediction system developers thus need an expanded set of diagnostic tools that will indicate which model development pathway will lead to rapid progress.

Let us consider a specific scientific challenge, diagnosing the sources of errors that are reflected in the ensemble initialization of T_{2m}. Consideration of this variable is of practical interest, for accurate T_{2m} forecasts directly help users (what will be tomorrow's high temperature?) and indirectly help them, for whether a thunderstorm develops often depends on T_{2m}. This variable is also commonly observed at surface weather stations. Unfortunately, 2 m above ground is commonly not a model level (~10 m is the lowest level in the ECMWF system), and so T_{2m} is instead an interpolated quantity. The fidelity of the interpolation often depends on stability and surface characteristics and their delicate interactions (ECMWF 2019a, section 3b). Thus, the diagnostics of interpolated T_{2m} provide indirect evidence about the character of the directly forecast model variables such as surface skin temperatures or the atmospheric model's first vertical level above ground.

Suppose then that we have found U-shaped rank histograms (Hamill 2001) of the initialized ensemble of T_{2m}. Were a high-quality, unbiased reference analysis of T_{2m} available with errors independent from those of the prediction system, it could potentially be compared against the ensemble's initial T_{2m} state estimates. Figure 1 illustrates how a reference T_{2m} analysis and error decomposition could help inform the categorization of errors and the model development process. If the differences between the operational initial states and the reference analysis were nonzero and consistent over time (systematic), changes to the underlying model may be warranted; the analysis bias was probably originating in part from biases in the underlying background state used in the data assimilation. Assuming that the remaining errors, the residual, were random in character and could be quantified, as many have proposed, the time-averaged spread of initial ensemble of analyses should be consistent with the time-averaged spread of the residual error after an incorporation of the effects of analysis uncertainty (Hamill 2001, Fig. 6., Saetra et al. 2004; Candille and Talagrand 2008; Weijs and van de Giesen 2011; Yamaguchi et al. 2016; Ben Bouallègue et al. 2020). If the time-averaged spread in the initial ensemble with analysis uncertainty was smaller than this residual component, then methodological changes to the ensemble initialization or stochastic physics procedures may be necessary to increase spread. The decomposition of error leveraging an independent analysis thus assists in determining which aspect of the prediction system should be improved, the underlying model or the ensemble methods. Performing the analysis with gridded data and not just observations permits an examination of spatial patterns of the errors. While only applied to initial conditions in this study and not to forecasts, the procedure is general and could be so applied.

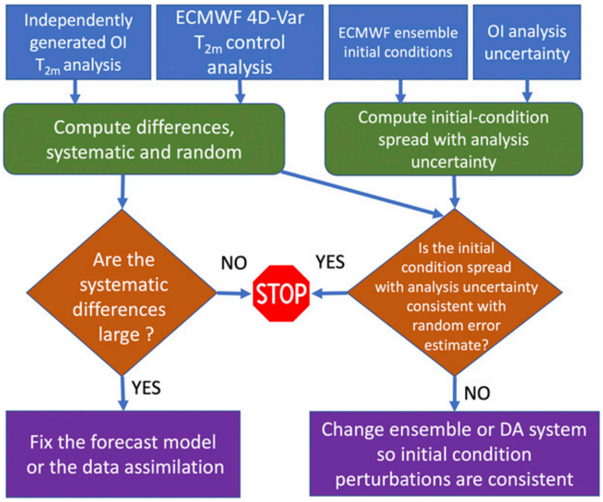

Independently generated OI T_{2m} analysis
ECMWF 4D-Var T_{2m} control analysis
ECMWF ensemble initial conditions
OI analysis uncertainty
Compute differences, systematic and random
Compute initial-condition spread with analysis uncertainty
Are the systematic differences large ?
NO
STOP
YES
Is the initial condition spread with analysis uncertainty consistent with random error estimate?
YES
NO
Fix the forecast model or the data assimilation
Change ensemble or DA system so initial condition perturbations are consistent

FIG. 1. The proposed diagnostic procedure for characterizing 2-m temperature analysis errors and what prediction-system changes may be necessary.

This error decomposition was applied to T_{2m} initialization in the ECMWF system in 2018 in this study. ECMWF has a somewhat complicated initialization procedure. ECMWF produces a T_{2m} analysis using optimal interpolation (ECMWF 2019b, chapter 9) and surface observations. However, this is not used directly in the atmospheric model initialization, but only indirectly to make increments to the soil moisture state. ECMWF's 2018 initial ensemble-state estimates of T_{2m} were not directly analyzed but were a vertically interpolated output of the atmospheric 4D-Var procedure. The 4D-Var procedure assimilated a wide variety of observations, but *not* surface T_{2m}. An independently produced T_{2m} analysis from surface observations not leveraging ECMWF background forecasts is thus proposed as a useful and independent dataset for evaluation of the ECMWF T_{2m} state ensemble estimates, with the acknowledged complication that T_{2m} is an interpolated quantity.

While we illustrate a particular error decomposition that leverages the independently generated T_{2m} analyses, we note that there is a rich literature of error decompositions and diagnostics related to weather prediction. Recent examples include Leutbecher (2010), Christensen (2015), Magnusson (2017), Rodwell et al. (2018), and many more as discussed in Wilks (2011, chapter 8). Still, the relatively simple approach here assuming a commonplace variance decomposition and Gaussianity will be demonstrated to provide some valuable insights into the characteristics of initialization of the ECMWF ensemble.

To apply the chosen diagnostic technique, we require high-quality, unbiased reference analyses of T_{2m}. Hamill and Scheuerer

(2020) described an optimal interpolation (OI; Gandin 1965; Daley 1991) procedure for creating a very high-resolution, accurate, unbiased, gridded statistical 1-h *forecast* of surface temperature over the contiguous United States (CONUS). With slight modification, this procedure was adapted to provide accurate analyses of 2-m temperature and its uncertainty, leveraging the OI procedure and a high-resolution, spatially detailed climatological background. A time series of these T_{2m} analyses and their uncertainty will be used in the diagnosis of ECMWF control T_{2m} initial-state estimate and the ensemble initialization of T_{2m}. Our hypothesis is that the diagnostic procedure will demonstrate issues both related to systematic error in the underlying prediction system and problems with adequate spread in the ensemble initialization.

Section 2 will provide details on the observation, climatology, and forecast data used in the study. The procedure used to generate the reference OI analyses is briefly described, with more complete detail in the online supplemental material (https://doi.org/10.1175/MWR-D-20-0119.s1). Section 2 will also describe the diagnostic procedure used to partition the initial ensemble errors into systematic and random components. Section 3 provides results, and section 4 discusses these and concludes. An appendix describes the procedure for estimating statistical significance of the analysis bias.

2. Data and methods

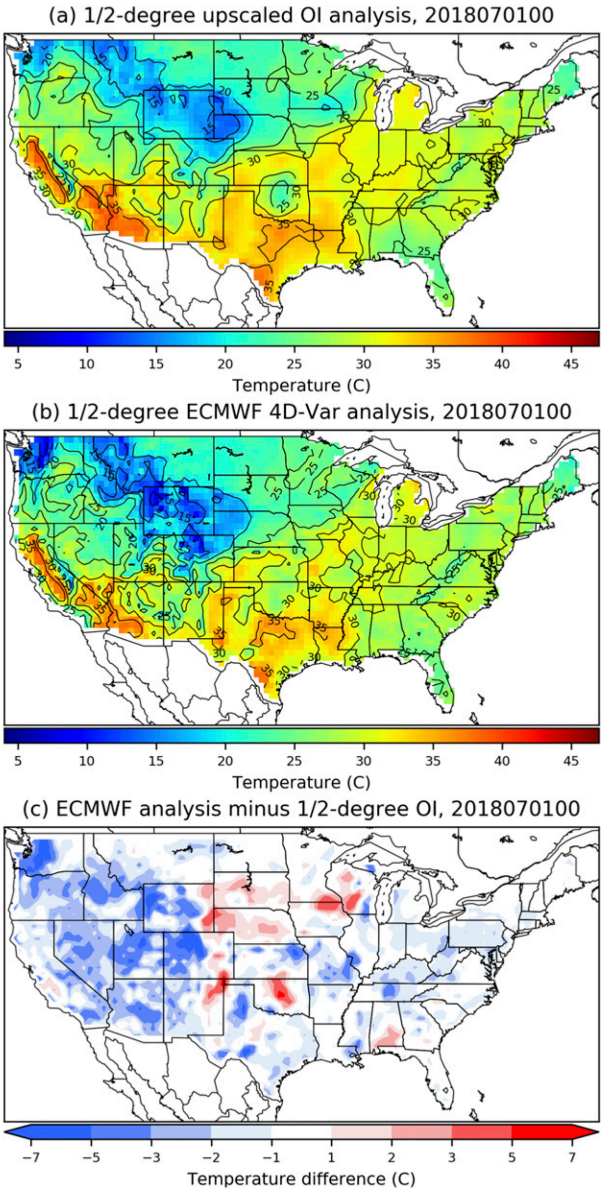
a. *Control initial state and ECMWF ensemble initial conditions*

ECMWF T_{2m} initial-condition data were downloaded from ECMWF's THORPEX Interactive Grand Global Ensemble (TIGGE; Bougeault et al. 2010; Swinbank et al. 2016) data portal on a 1/2° grid surrounding the contiguous United States (CONUS). A control state and 50-member ensemble initial conditions centered around the control were downloaded for every day in 2018. The ensemble initial conditions were created with perturbations generated from an ensemble of data assimilations (EDA; Bonavita et al. 2011) using a lower-resolution, perturbed-observation version of 4D-Var combined with singular-vector perturbations (Molteni et al. 1996). The EDA conducts parallel 4D-Var cycles, using different background forecasts for every member and perturbing the observations with independent random noise for each ensemble member. The added observation perturbations are random draws consistent with the observation-error covariance matrix.

The ECMWF prediction system changed during the year; prior to 5 June 2018, cycle 43r3 of the prediction system was used. At and after that date, cycle 45r1 was used. Documentation on the ECMWF prediction system versions can be found at https://www.ecmwf.int/en/publications/ifs-documentation. We did not examine the characteristics of the T_{2m} analysis time series to see if there were any discontinuities resulting from the change in model versions.

b. *Data and procedure used to generate reference OI analyses*

The observation dataset used in the generation of the OI analyses of T_{2m} was the National Center for Atmospheric Research (NCAR) dataset 472.0, an archive of quality-controlled hourly surface observations over North America. Data were originally synthesized and quality controlled at the U.S. National Weather Service Meteorological Development Laboratory. These data are available at https://rda.ucar.edu/datasets/ds472.0/. Surface temperatures were used at 0000 UTC for every day in 2018. The authors chose to further limit use of surface temperatures in this dataset to only those observation sites where data was available at 97% or more of the hours, days, and years in the analysis period. With this availability criterion, 1118 station locations were available in the area of study, the CONUS.


(a) 1/2-degree upscaled OI analysis, 2018070100

(b) 1/2-degree ECMWF 4D-Var analysis, 2018070100

(c) ECMWF analysis minus 1/2-degree OI, 2018070100

FIG. 2. (a) Upscaled 1/2° T_{2m} OI analysis for 0000 UTC 1 Jul 2018. (b) Corresponding ECMWF control analysis of T_{2m} produced by the 4D-Var. (c) ECMWF–OI analysis.

The procedure used to generate the high-resolution reference OI analyses of $T_{2\rm m}$ and used to estimate their uncertainty is described in the online supplemental material, along with an example (https://doi.org/10.1175/MWR-D-20-0119.s1). The procedure was inspired by and partially described in Hamill and Scheuerer (2020). The climatologies that provide the data assimilation background forecasts were derived from the Parameter-elevation Relationships on Independent Slopes Model (PRISM; Daly et al. 2008 and references therein). Observations were then used to produce a gridded set of analysis increments, the analyzed deviations from climatology. These were then added back to the 0000 UTC climatology to generate 0000 UTC analyzed fields, one for each day in 2018. The procedure for development of the 1/2° analysis uncertainty estimates are also described in the online supplemental material (https://doi.org/10.1175/MWR-D-20-0119.s1). This procedure is somewhat more complex, involving a determination of the high-resolution grid analysis-error covariances and an estimation of how much these are reduced in the upscaling to the 1/2° grid.

Comparison against ECMWF ensemble initial states in this study required upscaling of the high-resolution OI analyses to 1/2°, the grid spacing of ECMWF data in the TIGGE archive. Figure 2a shows an example of the "budget" upscaling (Accadia et al. 2003) of a sample OI analysis to the 1/2° grid of the ECMWF data. Figure 2b presents the control ECMWF initial state at the same time masked to the CONUS, and Fig. 2c shows the difference between the two. The ECMWF analysis for this date was colder over more grid points than it was warmer. Whether these differences were systematic will be evaluated through the use of an error decomposition applied at each grid point.

c. The proposed error decomposition and diagnostics

Let $T_t^{\rm true}$ represent the unknown discretized true state of $T_{2\rm m}$ at a grid point in question at date t on the 1/2° grid. Similarly, let $T_t^{\rm 4D}$ represent the ECMWF control 4D-Var atmospheric analyzed state estimate of $T_{2\rm m}$, and let $T_t^{\rm OI}$ represent the upscaled reference OI analyzed state, the reference estimate of the truth. We define $T_t^{\rm OI}$ as

$$T_t^{\rm OI} = T_t^{\rm true} + \epsilon_t; \quad \epsilon_t \sim N(0, \sigma_a^2). \qquad (1)$$

That is, the OI estimate was the true state plus an error, assumed to have zero mean and analysis-error variance σ_a^2 Justification for the unbiased character of the OI analysis can be found in Hamill and Scheuerer (2020). The difference of the 4D-Var analysis from the OI was then

$$\delta_t = T_t^{\rm 4D} - T_t^{\rm OI} = T_t^{\rm 4D} - T_t^{\rm true} - \epsilon_t. \qquad (2)$$

The assumption was that if the errors were consistent from day to day, they represented an error that was systematic. We thus assumed δ_t can be decomposed into a systematic component $\overline{\delta}_t$ that was slowly varying and a residual δ_t' that comprised the remaining error:

$$\delta_t = \overline{\delta}_t + \delta_t'. \qquad (3)$$

The systematic component for a particular date was then estimated with a temporal kernel smoother (Hastie and Tibshirani 1990, section 2.6; Rasmussen and Williams 2006). A Gaussian kernel was chosen that has a kernel radius b of 10 days:

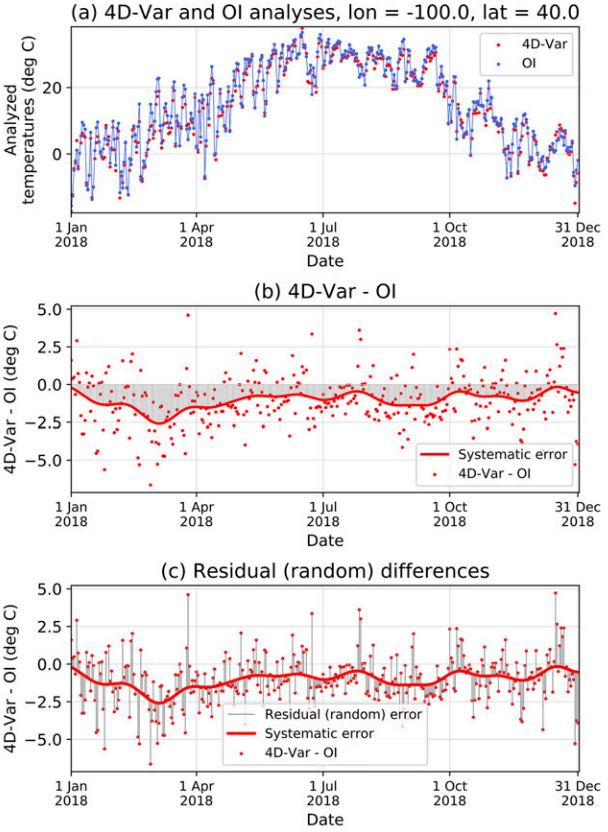


FIG. 3. (a) ECMWF 4D-Var and OI analysis time series in 2018 for a point along the Nebraska–Kansas border in the central United States. Blue lines connect the OI analyses and highlight the synoptic variability. (b) 4D-Var minus OI differences. Solid red line indicates the systematic error estimate. Gray shading emphasizes its magnitude. (c) As in (b), but gray lines now emphasize the residual, or random error.

$$K_\sigma(\delta_{t*}, \delta_t) = \exp\left[-\frac{(\delta_{t*}, \delta_t)^2}{2b^2}\right]. \qquad (4)$$

We then estimate $\overline{\delta}_t$ as

$$\overline{\delta}_t = \frac{\sum_{t*=1}^{365} K_\sigma(\delta_{t*}, \delta_t)\delta_{t*}}{\sum_{t*=1}^{365} K_\sigma(\delta_{t*}, \delta_t)}. \qquad (5)$$

The b in Eq. (4) was chosen through a cross-validation approach. For each day, temperatures for the days $t - 1$, t, and $t + 1$ were withheld, and a variety of possible b were evaluated. The cross-validated fit to the withheld data at time t were then evaluated for each choice of b. Setting $b = 10$ days provided an approximate best fit (lowest root-mean-square error) when averaged over all grid points across the CONUS and all days during 2018, though 20 and 30 days kernel radius (and longer)

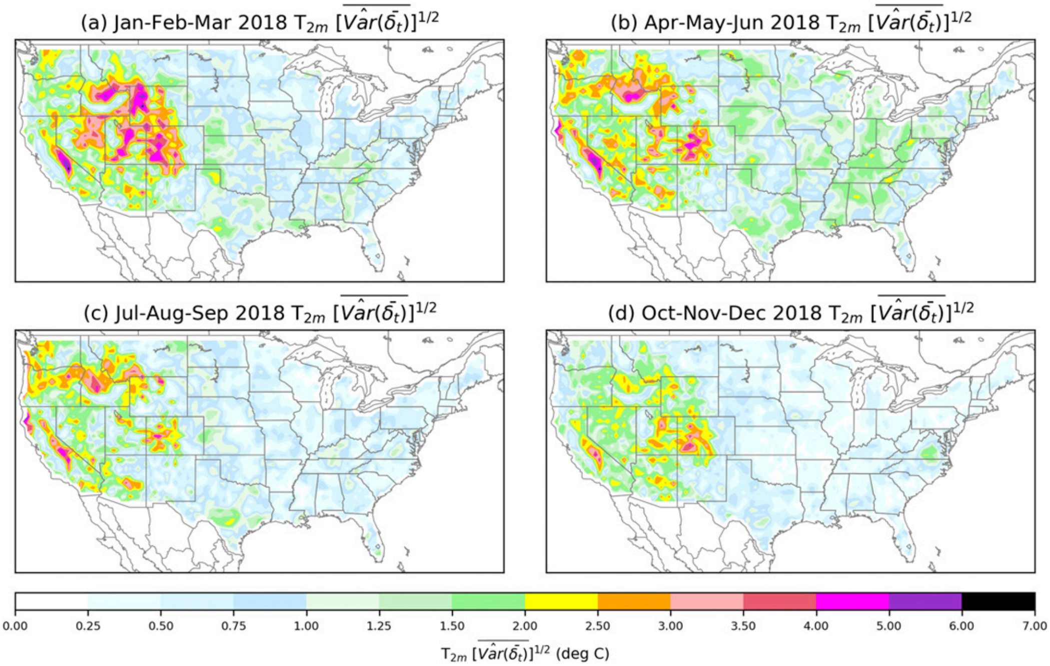

(a) Jan-Feb-Mar 2018 T_{2m} $[\overline{\hat{Var}(\bar{\delta}_t)}]^{1/2}$
(b) Apr-May-Jun 2018 T_{2m} $[\overline{\hat{Var}(\bar{\delta}_t)}]^{1/2}$
(c) Jul-Aug-Sep 2018 T_{2m} $[\overline{\hat{Var}(\bar{\delta}_t)}]^{1/2}$
(d) Oct-Nov-Dec 2018 T_{2m} $[\overline{\hat{Var}(\bar{\delta}_t)}]^{1/2}$
0.00
0.25
0.50
0.75
1.00
1.25
1.50
2.00
2.50
3.00
3.50
4.00
5.00
6.00
7.00
T_{2m} $[\overline{\hat{Var}(\bar{\delta}_t)}]^{1/2}$ (deg C)

FIG. 4. Estimate of the magnitude of $T_{2\mathrm{m}}$ systematic error for (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018.

were nearly equally as accurate. The surrounding days were withheld given the nonzero autocorrelation of presumed analysis error; this was a consequence of cycled data assimilation, where an error in the background was somewhat reflected in the 4D-Var analysis and then in the subsequent background forecast.

Let us now consider the variance decomposition of the 4D-Var minus OI time series:

$$\mathrm{Var}(\delta_t) = \mathrm{Var}(\bar{\delta}_t + \delta'_t) = \mathrm{Var}(\bar{\delta}_t) + \mathrm{Var}(\delta'_t) + 2\mathrm{cov}(\bar{\delta}_t\delta'_t). \quad (6)$$

For a given grid point, the left-hand side can be computed directly from the time series of differences in Eq. (2). The first term on the right-hand side $\bar{\delta}_t$ was the variance associated with the time-dependent estimate of the systematic error from the kernel smoother. The second term was the variance associated with the residual error δ'_t, and the third term was the covariance between the two. Ideally, these first and third terms would be near zero. When they are large, it may be instructive to examine the associated bias estimate and use this as a gateway to further investigations of their source(s).

An illustrative sample of the underlying data used to generate the terms in Eq. (6) are presented for a grid point in the central United States along the Nebraska–Kansas state border in Fig. 3. Time series of both the 4D-Var (red dots) and the OI (blue dots and lines) are shown in Fig. 3a. The solid blue lines in this panel illustrate the day-to-day weather variability. In Fig. 3b, the red dots illustrate the 4D-Var minus OI analysis differences, removing the common day-to-day weather-related variability. The thick red line in this panel provides the time-varying estimate of $\bar{\delta}_t$, which is below zero (i.e., a 4D-Var cold bias) to a greater extent in the winter and spring. The bias that presumably could be reduced through changes to the prediction system is highlighted with the gray shading. The individual black lines in Fig. 3c show the 4D-Var minus OI analysis differences from the time-varying mean error, the estimate of the residual (random) component. Variances associated with this term, it is asserted, should be consistent with the sum of ensemble initial-condition variance and the OI analysis-error variance, presumed to be independent, as no $T_{2\mathrm{m}}$ observations were assimilated during the 4D-Var assimilation procedure. In the results section, plots will be presented of the variance components averaged over 3-month periods. An overline $\overline{(\cdot)}$ is used to express time-averaged variance statistics, and hatted functions $\hat{}$ indicate sample statistics. For ease of interpretation, spreads (standard deviations) will be plotted instead of variances, i.e., $\left[\overline{\hat{\mathrm{Var}}\,(\cdot)}\right]^{1/2}$.

A primary question of importance to the ensemble prediction system developer is whether the initial spread in the ensemble is of an appropriate magnitude. The ith member of a 50-member $T_{2\mathrm{m}}$ ensemble initial condition at time t and at a grid point will be referred to as $T_t^i(\mathrm{ens})$ The sample variance of that ensemble over all members will be referred to as $\hat{\mathrm{Var}}[\mathbf{T}_t(\mathrm{ens})]$ We propose that the time average of spreads in an ensemble after adding a term that incorporates analysis uncertainty should be matched in magnitude by the time average of the residual component of the analysis error: $\left\{\overline{\hat{\mathrm{Var}}[\mathbf{T}_t(\mathrm{ens})] + \hat{\mathrm{Var}}_t^{\mathrm{OI}}}\right\}^{1/2} \sim \left[\overline{\hat{\mathrm{Var}}(\delta'_t)}\right]^{1/2}$ This consistency can be readily checked presuming we have an estimate of the analysis-error variance. The procedure for estimating this was rather involved and again is discussed in the online supplemental material (https://doi.org/10.1175/MWR-D-20-0119.s1).

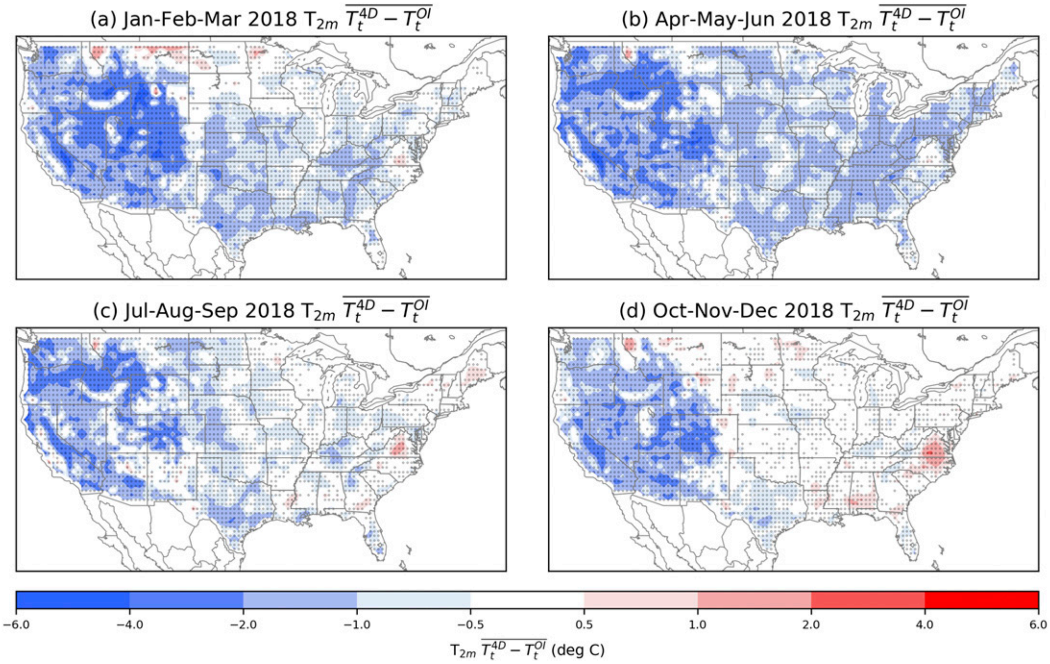

(a) Jan-Feb-Mar 2018 T_{2m} $\overline{T_t^{4D} - T_t^{OI}}$
(b) Apr-May-Jun 2018 T_{2m} $\overline{T_t^{4D} - T_t^{OI}}$
(c) Jul-Aug-Sep 2018 T_{2m} $\overline{T_t^{4D} - T_t^{OI}}$
(d) Oct-Nov-Dec 2018 T_{2m} $\overline{T_t^{4D} - T_t^{OI}}$
−6.0
−4.0
−2.0
−1.0
−0.5
0.5
1.0
2.0
4.0
6.0
T_{2m} $\overline{T_t^{4D} - T_t^{OI}}$ (deg C)

FIG. 5. Time-mean 4D-Var bias by season with respect to the upscaled OI analysis. (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018. Points where the departure from 0 is determined to be statistically significant using the procedure described in the appendix are dotted.

Our error decomposition bears similarity to others in the literature such as Rodwell et al. (2018; R18). But there are differences. In this study, analysis uncertainty replaces the observation uncertainty of R18, and the residuals of R18 are implicitly estimated through the comparison of spread ratios.

3. Results

First we consider the magnitude of systematic error and bias. Figure 4 shows gridded estimates of systematic-error spread $\left[\overline{\hat{\mathrm{Var}}(\overline{\delta}_t)}\right]^{1/2}$ for each season. The most notable feature of Fig. 4 is that the systematic-error estimates were larger in the western United States and were somewhat larger in the winter and spring than they were in the fall and summer. These were manifested in temperatures that were cooler on average in the 4D-Var system than they were in the OI (Fig. 5). The differences were significant for a large number of grid points.

Perhaps numerical issues with the vertical interpolation to the 2-m level were partially responsible for the systematic cold bias of 2-m temperatures, not accounting for the variety of possible behaviors in different atmospheric conditions. Another possible source was that differences in grid elevation may increase the systematic analysis-error variance in the western United States. This possibility is examined in Fig. 6, which compared terrain elevations of the 4-km OI computational grid upscaled to the ECMWF grid, and their differences. While there were small-scale differences when the ECMWF data were interpolated to the OI analysis grid spacing, when the OI analysis terrain were budget upscaled to the 1/2° spacing at which the ECMWF data was saved in the TIGGE database, the differences were small and apparently random in character; there did not appear to be any systematic elevation difference. The larger systematic differences in the cool-season that were especially prominent in mountainous regions suggests another possible hypothesis: perhaps the fractional snow cover or snow depth were misestimated, causing systematic biases in the background $T_{2\mathrm{m}}$ used in the 4D-Var system. These were not evaluated here but may be a subject of further research at ECMWF.

Now consider whether the initial-condition spread of $T_{2\mathrm{m}}$ was accurately estimated. Figure 7 presents the time-averaged, random (residual) spread $\left[\overline{\hat{\mathrm{Var}}(\overline{\delta}_t)}\right]^{1/2}$, which should be matched by the initial-condition spread after accounting for OI analysis uncertainty, $\left\{\overline{\hat{\mathrm{Var}}[\mathbf{T}_t(\mathrm{ens})] + \hat{\mathrm{Var}}_t^{\mathrm{OI}}}\right\}^{1/2}$. The time average of the residual spread was generally larger in the mountains of the western United States than in the eastern United States, though the region with largest residual spread shifts with season from the northern Rockies in the fall and winter to the eastern Rockies and western Great Plains in the spring and the southern Rockies in the summer.

The construction of $\left\{\overline{\hat{\mathrm{Var}}[\mathbf{T}_t(\mathrm{ens})] + \hat{\mathrm{Var}}_t^{\mathrm{OI}}}\right\}^{1/2}$ is now considered. Figure 8 shows the time average of ensemble initial-condition spreads for each season, $\left\{\overline{\hat{\mathrm{Var}}[T_t(\mathrm{ens})]}\right\}^{1/2}$. The spreads by themselves were much lower in magnitude than the time average of the residual spread in the previous figure.

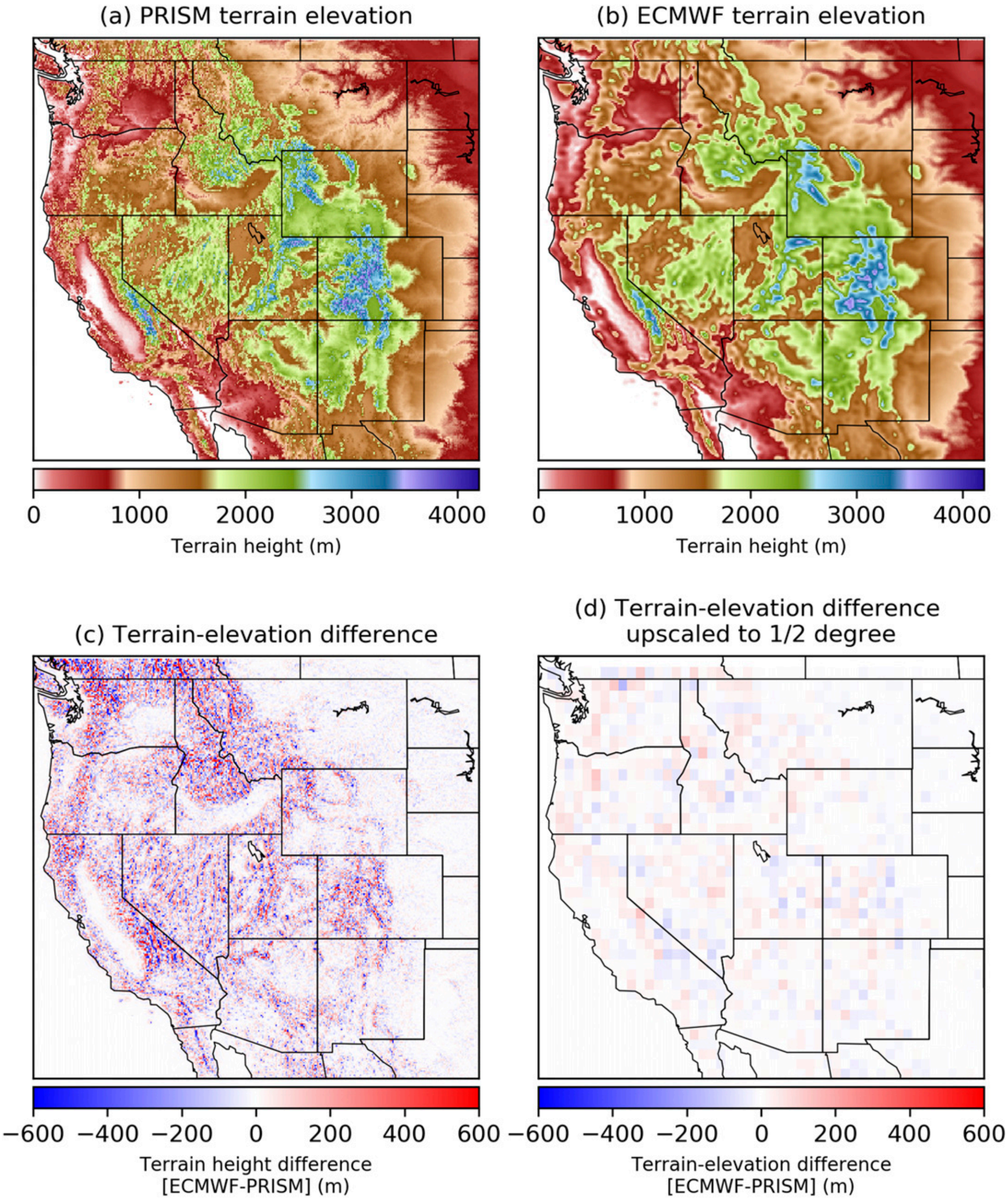

(a) PRISM terrain elevation
0
1000
2000
3000
4000
Terrain height (m)
(b) ECMWF terrain elevation
0
1000
2000
3000
4000
Terrain height (m)
(c) Terrain-elevation difference
−600
−400
−200
0
200
400
600
Terrain height difference
[ECMWF-PRISM] (m)
(d) Terrain-elevation difference
upscaled to 1/2 degree
−600
−400
−200
0
200
400
600
Terrain-elevation difference
[ECMWF-PRISM] (m)

FIG. 6. (a) PRISM terrain elevation. (b) ECMWF terrain elevation. (c) Elevation difference. (d) Elevation difference upscaled to 1/2° grid.

The ensemble spreads were generally larger in the mountainous western United States, but were larger in the summer than in the winter. The estimates of the OI analysis uncertainty, $\left[\overline{\mathrm{V\hat{a}r}_t^{\mathrm{OI}}}\right]^{1/2}$ are shown in Fig. O12 of the online appendix (https://doi.org/10.1175/MWR-D-20-0119.s1). These exhibit uncertainty maxima in the northern and central Rockies and in the Cascade range of Washington state and Oregon. The larger uncertainty there was consistent with the sparser observation network in these areas and the more limited spreading of observation data due to the background-error covariance model, which limits spreading of data influence from valley locations to high terrain. The sum of the two terms are provided in Fig. 9; it is this that should be consistent in magnitude with the residual spread of Fig. 7. The sum in Fig. 9 again exhibited maxima in the mountainous western United States, with a consistent peak in the Yellowstone area of northwest Wyoming. Generally, the areas with larger analysis uncertainty follow station density (Fig. 2b) and complexity of terrain elevation. Unfortunately, the large analysis uncertainty in the western United States made determination of whether the ensemble was under versus overspread

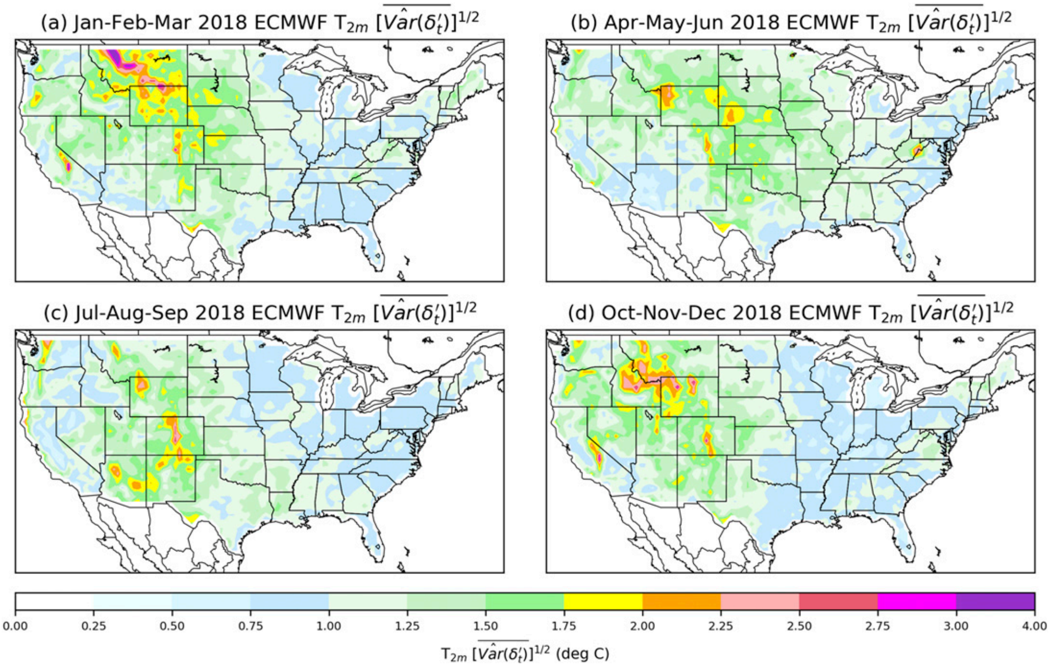

(a) Jan-Feb-Mar 2018 ECMWF T_{2m} $[\overline{\hat{Var}(\delta'_t)}]^{1/2}$
(b) Apr-May-Jun 2018 ECMWF T_{2m} $[\overline{\hat{Var}(\delta'_t)}]^{1/2}$
(c) Jul-Aug-Sep 2018 ECMWF T_{2m} $[\overline{\hat{Var}(\delta'_t)}]^{1/2}$
(d) Oct-Nov-Dec 2018 ECMWF T_{2m} $[\overline{\hat{Var}(\delta'_t)}]^{1/2}$
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
4.00
T_{2m} $[\overline{\hat{Var}(\delta'_t)}]^{1/2}$ (deg C)

FIG. 7. Estimate of time-averaged random (residual) spread: (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018.

challenging, for initial ensemble spread in this area was much lower than the analysis uncertainty. That is, the sum of the two mostly reflected the contributions of the analysis uncertainty.

To facilitate interpretation where analysis uncertainty was lower, the ratios of the data in Fig. 9 (numerator) versus Fig. 7 (denominator) are plotted in Fig. 10. This helps interpret where the initial ensemble may have been under or overspread

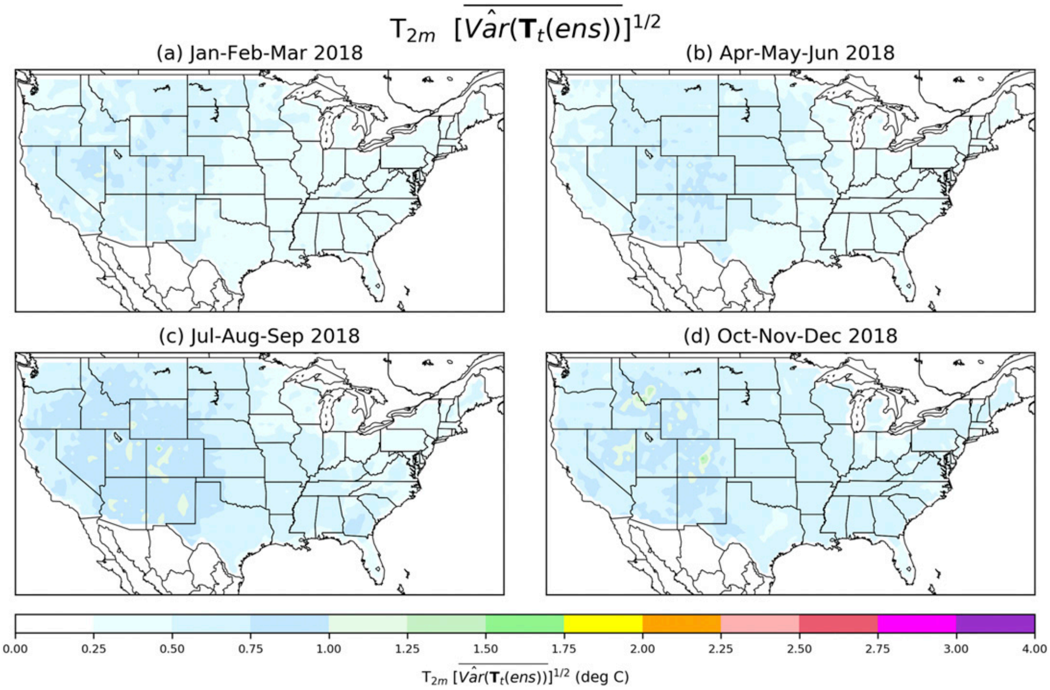


FIG. 8. Estimate of $T_{2\mathrm{m}}$ time-averaged ensemble initial spread: (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018.

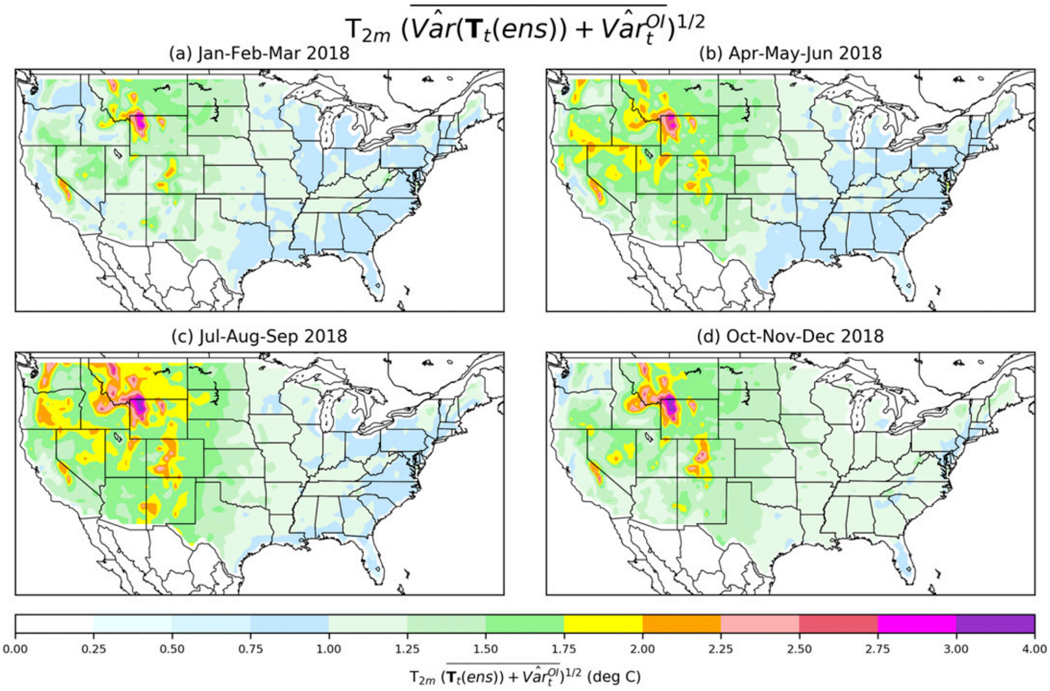

T_{2m} $(\overline{\hat{Var}(\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}})^{1/2}$
(a) Jan-Feb-Mar 2018
(b) Apr-May-Jun 2018
(c) Jul-Aug-Sep 2018
(d) Oct-Nov-Dec 2018
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00
4.00
T_{2m} $(\overline{\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}})^{1/2}$ (deg C)

FIG. 9. Estimate of T_{2m} time-averaged ensemble initial spread (Fig. 8) plus OI analysis uncertainty (online appendix in the supplemental material, Fig. O12; https://doi.org/10.1175/MWR-D-20-0119.s1), calculated as the square root of the sum of the squares: (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018.

(blue versus red color). In the central and eastern United States, the initial spread estimate appeared more reasonable in magnitude with ratios near 1.0, though the diagnostic suggested underspread ensembles in April–June and overspread in October–December. Again, interpretation in the western United States was complicated by the large analysis uncertainty. The small signal (ensemble initial spread) was overwhelmed by the larger noise (analysis uncertainty). This could be ameliorated through assimilation of a denser network of observations. This analysis procedure used $O(10^3)$ observations; the PRISM procedure used $O(10^4)$. Suppose the OI analysis-error variances were uniformly decreased everywhere by a factor of 3 through assimilation of a more dense network of T_{2m} observations, changing the sum in Fig. 9.[1] Since these T_{2m} observations were not assimilated in the 2018 version of ECMWF's 4D-Var with its significant bias, the denominator was assumed to remain relatively unchanged. The revised ratio is plotted in Fig. 11, which shows that the ensemble with the assimilation of more dense observations was diagnosed as being underspread in most regions. It should be noted that this diagnostic does not account for the possibility of smaller residual errors in Fig. 7 resulting from a more accurate analysis. Hence the assertion that the ECMWF T_{2m} ensemble might be underspread would need to be checked through the actual OI assimilation of a denser network of observations; this was not performed here.

Were this diagnostic applied to longer-lead forecasts rather than initial conditions, then the ensemble spread would be larger, providing proportionately more signal to the noise (analysis uncertainty). Hence, the existing T_{2m} analyses may be better suited to diagnosis of medium-range forecast errors than initial-condition errors.

4. Discussion and conclusions

The purpose of this article was to demonstrate application of a simple error decomposition to initial conditions for ensemble weather–climate prediction, enabling some quantification of how much of the errors were systematic and how much were random. The decomposition was applied to the (interpolated) 2-m temperature (T_{2m}) initialization of the ECMWF ensemble in 2018, leveraging an independently generated optimal interpolation (OI) procedure of T_{2m} to provide the verification analyses and the estimates of analysis uncertainty. Given that the OI procedure draws closely to the observations, a direct comparison against the observations at these sites would produce similar results. The error decomposition showed that there were statistically significant systematic errors in T_{2m} initialization across the CONUS. Interpretation of

[1] The uniform factor-of-3 reduction is a simplification of what may happen in practice, where analysis-error variances will commonly decrease by a greater amount through the introduction of new observations in data-sparse regions compared to data-rich regions. See Morss et al. (2001, their Fig. 4).

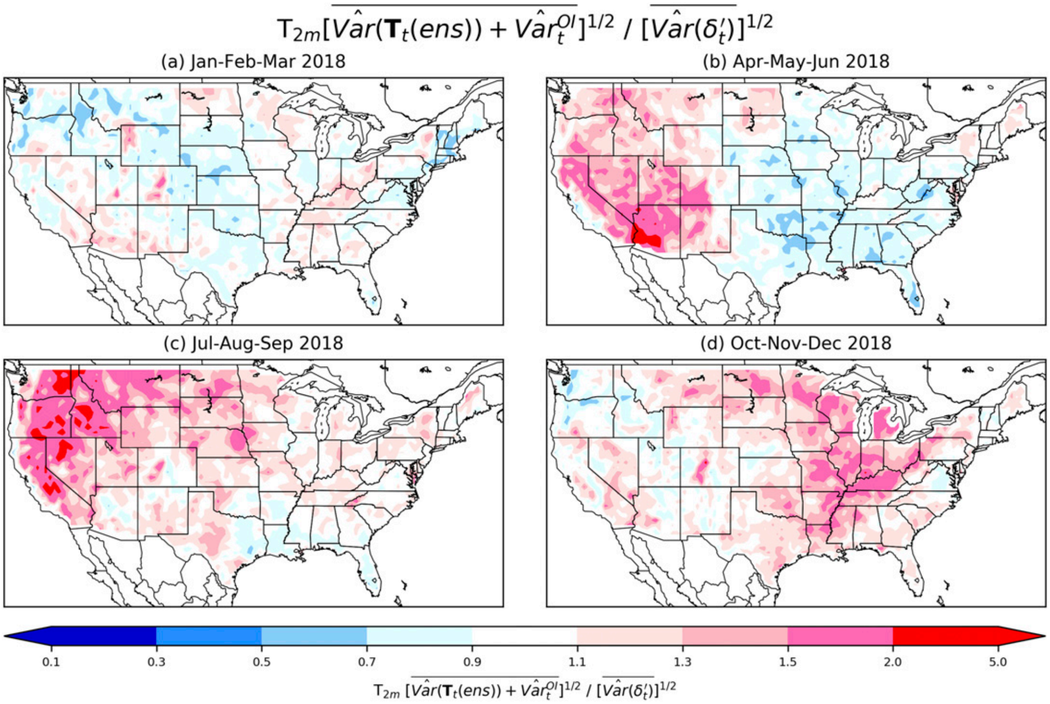

$T_{2m}[\overline{\hat{Var}(\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}}]^{1/2} / [\overline{\hat{Var}(\delta_t')}]^{1/2}$
(a) Jan-Feb-Mar 2018
(b) Apr-May-Jun 2018
(c) Jul-Aug-Sep 2018
(d) Oct-Nov-Dec 2018
0.1
0.3
0.5
0.7
0.9
1.1
1.3
1.5
2.0
5.0
$T_{2m}[\overline{\hat{Var}(\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}}]^{1/2} / [\overline{\hat{Var}(\delta_t')}]^{1/2}$

FIG. 10. Time-averaged ensemble initial spread plus OI analysis uncertainty (Fig. 9), (see online appendix; https://doi.org/10.1175/MWR-D-20-0119.s1) divided by the time-averaged random (residual) spread (Fig. 7). This provides an estimate of whether the initial ensemble is underspread (ratio < 1.0) or overspread (ratio > 1.0): (a) January–March, (b) April–June, (c) July–September, and (d) October–December 2018.

initial-condition spread character was less definite, in that the substantial OI analysis uncertainty made interpretation of the initial spread characteristic ambiguous. Overall, the initial hypothesis of biased initial ensemble T_{2m} states was confirmed, but the hypothesized underspread characteristic could not be successfully determined using the OI analyses and their comparatively large analysis uncertainty. In future work, this diagnosis could be improved by assimilating more observations to reduce that uncertainty, and the diagnostic could still be applied to medium-range ensemble forecasts, where forecast spread is larger than the analysis uncertainty.

The systematic errors for 0000 UTC showed up as a cold bias, one that was larger in the western and central United States during winter and spring. The systematic errors may necessitate changes to the underlying deterministic prediction system, and a deficiency in the initial spread may necessitate changes to the method of construction of the initial ensemble. A necessary next step is further exploratory data analysis combined with the insight of prediction system developers to identify the key underlying deficiencies. This is a challenging endeavor, for "*there is hardly any component of the atmospheric and land-surface modeling system that does not have an influence on* T_{2m} *errors*" (M. Bonavita 2020, ECMWF, personal communication). A subsequent exploratory data analysis to determine the ultimate sources of these errors was not possible in this article. Still, we suggest some possible directions, many of which are already being actively explored at ECMWF. A comparatively simple explanation may be the procedure for vertical interpolation to the T_{2m} level. For the 137-level version of the data assimilation system used during 2018 (https://www.ecmwf.int/en/forecasts/documentation-and-support/137-model-levels), again the lowest model level was at approximately 10 m above ground. The temperature analysis at the 2-m level is computed as an interpolation between the ground surface temperature and this lowest model level. Lapse rates between these two levels can vary greatly in space and time, with superadiabatic lapse rates in calm, clear daytime conditions to extreme inversions at night. The ECMWF procedure for interpolation to the 2-m level is discussed in ECMWF (2019a, section 3.10.3) and could be examined as a candidate for improvement.

Beyond possible issues with vertical interpolation between model levels, the land T_{2m} is strongly sensitive to the land surface energy partitioning. Incoming fluxes of net downward longwave and shortwave energy are balanced by surface sensible and latent heat fluxes and a ground heat flux. Improvement of initialization and modeling of the land surface energy balance is an active area of research at ECMWF (e.g., Haiden and Trentmann 2016; Hogan et al. 2017; Orth et al. 2017; Haiden et al. 2018b,c; Fairbairn et al. 2019; Munoz-Sabater et al. 2019). There are many opportunities for model systematic errors to deleteriously affect the surface energy balance and consequently the estimation of T_{2m}. These may include a misestimation of downward shortwave radiative fluxes through misestimation of cloud fraction and optical

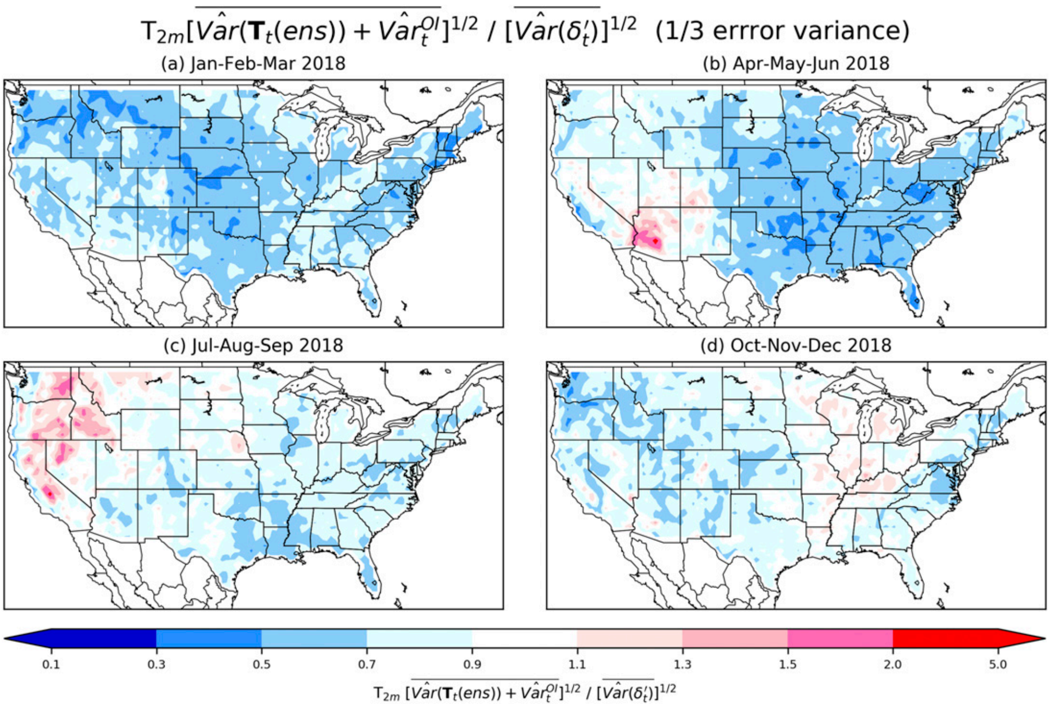

$T_{2m}[\overline{\hat{Var}(\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}}]^{1/2}$ / $[\overline{\hat{Var}(\delta_t^i)}]^{1/2}$ (1/3 errror variance)
(a) Jan-Feb-Mar 2018
(b) Apr-May-Jun 2018
(c) Jul-Aug-Sep 2018
(d) Oct-Nov-Dec 2018
0.1
0.3
0.5
0.7
0.9
1.1
1.3
1.5
2.0
5.0
T_{2m} $[\overline{\hat{Var}(\mathbf{T}_t(ens)) + \hat{Var}_t^{OI}}]^{1/2}$ / $[\overline{\hat{Var}(\delta_t^i)}]^{1/2}$

FIG. 11. As in Fig. 10, but assuming the analysis-error variance has been decreased by a factor of 3 through the assimilation of more observations while residual spread (Fig. 7) remains unchanged.

depth, which in turn can be strongly sensitive to the methods used to parameterize deep and shallow convection and cloud microphysics. Misestimation of boundary and surface-layer mixing also affects the surface sensible and latent heat fluxes. In the daytime, unduly calm winds and drier soils will result in larger ground heat storage and warmer T_{2m}. There are also many challenges associated with correctly modeling evaporative fluxes from the soil and canopy, including misspecification of model and physiographic parameters such as those related to soil hydraulic conductivity, stomatal resistance, leaf area index, the parameterization of fractionally snow-covered soil, albedo, roughness lengths, urban fraction, vegetative fraction, and more.

Systematic forecast errors of T_{2m} may in turn be related to errors in the land surface initial state, including its soil moisture and temperature and snow cover. If initial snow depth and fractional snow cover are overestimated, it is likely that the forecast temperatures used as background for the next data assimilation cycle will be biased too low and remain low in the absence of adjustment to new observations. As temperatures had a cold bias in the mountainous western United States during winter and spring, the accuracy of initial snow estimates (ECMWF 2019b, section 9.3) could be investigated further.

The diagnostics also indicated that even with the large OI analysis uncertainty, sometimes the initial T_{2m} ensemble spread was somewhat too small. Changes that may result in greater spread in the ECMWF system might include: (i) inducing a greater variety in background soil temperatures and soil moistures, perhaps through application of stochastic physics in the soil state. These affect the surface-energy balance among ensembles and thus T_{2m} variety. (ii) Increasing the variety of snow amount and fractional snow cover in initial state, be it through stochastic physics that create a greater variety of ensemble precipitation and surface-temperature forcings, or perturbations to their initial state. (iii) Applying techniques that introduce a greater variety of downward solar radiation to the surface; perhaps parameters in the SPPT stochastic physics (Leutbecher et al. 2017) can be changed, or more physically based stochastic parameterizations of deep convection and/or microphysics developed. These drive variety in energy partitions at the surface. (iv) Update the methodology of perturbing the observations in the ensembles of data assimilations procedure. If correlations of errors are not correctly specified in the observation-error covariance matrix **R** used in the 4D-Var (possibly observation errors are assumed to be more independent than they are), this will result in an ensemble with lower posterior spread.

In this application, the diagnostics leveraged surface-temperature analyses that were generated independently through a procedure with low bias but with higher-than-desirable analysis uncertainty in the western United States. Assimilation of more observations would reduce the analysis uncertainty, making this OI procedure possibly more relevant for diagnosing initial-condition uncertainty characteristics. The procedure is general, though, and its limitations would be less severe when applied to, say, medium-range forecasts where the forecast spread is larger and the OI analysis uncertainty is comparatively smaller.

Could the OI analysis be replaced with one generated internally at ECMWF? It is possible that another validation dataset such as the T_{2m} analysis produced as part of the global ERA5 dataset (Hersbach et al. 2019) could be used. However, this and other similar products leverage a background forecast created by a numerical model, one which may have biases that are nontrivial and related to the biases in the operational model initialization. Further research is needed to determine the suitability of internally generated ECMWF reanalyses as the reference for the error decomposition.

Acknowledgments. Maria Gehne (PSL and U. Colorado/CIRES) is acknowledged for her research (Gehne et al. 2019) that stimulated the development of this error decomposition. Matt Switanek (U. Colorado/CIRES and PSD) provided the python code that implements the spatial statistical significance testing of bias, following Wilks (2016). Thomas Haiden and Martin Leutbecher (ECMWF) are gratefully acknowledged for their constructive criticism during the research process. Massimo Bonavita (ECMWF) and two anonymous reviewers are thanked for their constructive suggestions for improving the manuscript. Funding for this project was provided through ESRL/PSL base funds and through an Office of Weather and Air Quality grant to CIRES (NA17OAR4320101).

APPENDIX

Statistical Significance Testing of Bias

The procedure for the statistical significance testing of gridded bias estimates follows best practices for multiple simultaneous statistical hypothesis tests, following Wilks (2016) and references therein. The procedure began with the determination of p values associated with a one-sample two-sided t test of bias. Because of the autocorrelation of bias estimates, an effective sample size was estimated at each grid point using the lag-1 autocorrelation estimate ρ determined with a Pearson product-moment correlation. Then, following Wilks (2011, Eq. 5.12) the effective sample size was computed as

$$n' \simeq \frac{1-\rho}{1+\rho}. \tag{A1}$$

This effective sample size was used in the computation of the variance estimate, the denominator of the t statistic:

$$t = \frac{\overline{x} - \mu_0}{\left[\widehat{\mathrm{Var}(\overline{x})}\right]^{1/2}}, \tag{A2}$$

where μ_0 was the null hypothesis value (here, zero), and $\overline{x}$ was the sample mean of the time series of δ_t, 4D-Var minus OI differences. The estimated variance of the sample mean $\widehat{\mathrm{Var}(\overline{x})}$ was computed using the effective sample size:

$$\widehat{\mathrm{Var}(\overline{x})} = \frac{s^2}{n'}, \tag{A3}$$

where s was the sample variance of the time series of 4D-Var minus OI differences.

The procedure then directly follows Wilks (2016) for computation of statistically significant grid points when controlling the false discovery rate at 10%.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*, **18**, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Ahlgrimm, M., R. M. Forbes, R. J. Hogan, and I. Sandu, 2018: Understanding global model systematic shortwave radiation errors in subtropical marine boundary layer cloud regimes. *J. Adv. Model. Earth Syst.*, **10**, 2042–2060, https://doi.org/10.1029/2018MS001346.

Beljaars, A., and Coauthors, 2018: The numerics of physical parametrization in the ECMWF model. *Front. Earth Sci.*, **6**, 137, https://doi.org/10.3389/feart.2018.00137.

Ben Bouallègue, Z., T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. *Mon. Wea. Rev.*, **148**, 2049–2062, https://doi.org/10.1175/MWR-D-19-0323.1.

Bonavita, M., L. Raynaud, and L. Isaksen, 2011: Estimating background-error variances with the ECMWF ensemble of data assimilations system: Some effects of ensemble size and day-to-day variability. *Quart. J. Roy. Meteor. Soc.*, **137**, 423–434, https://doi.org/10.1002/qj.756.

——, E. Hólm, L. Isaksen, and M. Fisher, 2016: The evolution of the ECMWF hybrid data assimilation system. *Quart. J. Roy. Meteor. Soc.*, **142**, 287–303, https://doi.org/10.1002/qj.2652.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Buehner, M., P. L. Houtekamer, C. Charette, H. L. Mitchell, and B. He, 2010a: Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: Description of single-observation experiments. *Mon. Wea. Rev.*, **138**, 1550–1566, https://doi.org/10.1175/2009MWR3157.1.

——, ——, ——, ——, and ——, 2010b: Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: One-month experiments with real observations. *Mon. Wea. Rev.*, **138**, 1567–1586, https://doi.org/10.1175/2009MWR3158.1.

Buizza, R., 2018: Ensemble forecasting and the need for calibration. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. S. Wilks, and J. W. Messner, Eds., Elsevier Press, 15–48.

——, 2019: Introduction to the special issue on "25 years of ensemble forecasting." *Quart. J. Roy. Meteor. Soc.*, **145**, 1–11, https://doi.org/10.1002/qj.3370.

——, M. Miller, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Candille, G., and O. Talagrand, 2008: Impact of observational error on the validation of ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **134**, 959–971, https://doi.org/10.1002/qj.268.

Christensen, H. M., 2015: Decomposition of a new proper score for verification of ensemble forecasts. *Mon. Wea. Rev.*, **143**, 1517–1532, https://doi.org/10.1175/MWR-D-14-00150.1.

——, J. Berner, D. Coleman, and T. N. Palmer, 2017: Stochastic parameterization and El Niño–Southern Oscillation. *J. Climate*, **30**, 17–38, https://doi.org/10.1175/JCLI-D-16-0122.1.

Daley, R., 1991: *Atmospheric Data Analysis*. Cambridge University Press, 472 pp.

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. A. Pasteris, 2008: Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *Int. J. Climatol.*, **28**, 2031–2064, https://doi.org/10.1002/joc.1688.

ECMWF, 2019a: IFS Documentation—Cy46r1 operational implementation 6 June 2019. Part IV: Physical processes. ECMWF, accessed 1 August 2020, https://www.ecmwf.int/en/elibrary/19308-part-iv-physical-processes.

——, 2019b: IFS Documentation—Cy46r1 operational implementation 6 June 2019. Part II: Data assimilation. ECMWF, accessed 1 August 2020, https://www.ecmwf.int/en/elibrary/19306-part-ii-data-assimilation.

Fairbairn, D., P. de Rosnay, and P. A. Browne, 2019: The new stand-along surface analysis at ECMWF: Implications for land–atmosphere DA coupling. *J. Hydrometeor.*, **20**, 2023–2042, https://doi.org/10.1175/JHM-D-19-0074.1.

Gandin, L. S., 1965: *Objective Analysis of Meteorological Fields*. Israel Program for Scientific Translation, 242 pp.

Gehne, M., T. M. Hamill, G. T. Bates, P. Pegion, and W. Kolczynski, 2019: Land surface parameter and state perturbations in the global ensemble forecast system. *Mon. Wea. Rev.*, **147**, 1319–1340, https://doi.org/10.1175/MWR-D-18-0057.1.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.*, **69B**, 243–268, https://doi.org/10.1111/j.1467-9868.2007.00587.x.

Haiden, T., and J. Trentmann, 2016: Verification of cloudiness and radiation forecasts in the greater Alpine region. *Meteor. Z.*, **25**, 3–15, https://doi.org/10.1127/metz/2015/0630.

——, M. Janousek, J.-R. Bidlot, R. Buizza, L. Ferranti, F. Prates, and F. Vitart, 2018a: Evaluation of ECMWF forecasts, including the 2018 upgrade. ECMWF Tech. Memo. 831, 54 pp., https://www.ecmwf.int/sites/default/files/elibrary/2018/18746-evaluation-ecmwf-forecasts-including-2018-upgrade.pdf.

——, and Coauthors, 2018b: Use of in situ surface observations at ECMWF. ECMWF Tech. Memo. 834, 28 pp., https://www.ecmwf.int/sites/default/files/elibrary/2018/18748-use-situ-surface-observations-ecmwf.pdf.

——, I. Sandu, G. Balsamo, G. Arduini, and A. Beljaarsm, 2018c: Addressing biases in near-surface forecasts. *ECMWF Newsletter*, No. 157, ECMWF, Reading, United Kingdom, 20–25, https://www.ecmwf.int/en/newsletter/157.

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 J. S. Whitaker, 2007: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. *Mon. Wea. Rev.*, **135**, 3273–3280, https://doi.org/10.1175/MWR3468.1.

——, and M. Scheuerer, 2020: Benchmarking the raw model-generated background forecast in rapidly updated surface temperature analyses. Part II: Gridded benchmark. *Mon. Wea. Rev.*, **148**, 701–717, https://doi.org/10.1175/MWR-D-19-0028.1.

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models*. Chapman & Hall, 335 pp.

Hersbach, H., and Coauthors, 2019: Global reanalysis: Goodbye ERA-Interim, hello ERA5. *ECMWF Newsletter*, No. 59, ECMWF, Reading, United Kingdom, 7–24, https://www.ecmwf.int/node/19027.

Hogan, R., and Coauthors, 2017: Radiation in numerical weather prediction. ECMWF Tech. Memo. 816, 51 pp., http://ecmwf.int/sites/default/files/elibrary/2017/17771-radiation-numerical-weather-prediction.pdf.

Houtekamer, P. L., and H. L. Mitchell, 1998: Data assimilation using an ensemble Kalman filter technique. *Mon. Wea. Rev.*, **126**, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2.

——, and ——, 2001: A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.*, **129**, 123–137, https://doi.org/10.1175/1520-0493(2001)129<0123:ASEKFF>2.0.CO;2.

——, and ——, 2005: Ensemble Kalman filtering. *Quart. J. Roy. Meteor. Soc.*, **131**, 3269–3289, https://doi.org/10.1256/qj.05.135.

Juricke, S., D. MacLeod, A. Weisheimer, L. Zanna, and T. N. Palmer, 2018: Seasonal to annual ocean forecasting skill and the role of model and observational uncertainty. *Quart. J. Roy. Meteor. Soc.*, **144**, 1947–1964, https://doi.org/10.1002/qj.3394.

Lavaysse, C., M. Carrera, S. Belair, N. Gagnon, R. Frenette, M. Charron, and M. K. Yau, 2013: Impact of surface parameter uncertainties within the Canadian regional ensemble prediction system. *Mon. Wea. Rev.*, **141**, 1506–1526, https://doi.org/10.1175/MWR-D-11-00354.1.

Leutbecher, M., 2010: Diagnosis of ensemble forecasting systems. *Proc. Seminar on Diagnosis of Forecasting and Data Assimilation Systems*, Reading, United Kingdom, ECMWF, 235–266.

——, and Coauthors, 2017: Stochastic representations of model uncertainties at ECMWF: State of the art and future vision. *Quart. J. Roy. Meteor. Soc.*, **143**, 2315–2339, https://doi.org/10.1002/qj.3094.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.*, **20**, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors. *Quart. J. Roy. Meteor. Soc.*, **143**, 2129–2142, https://doi.org/10.1002/qj.3072.

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.

Morss, R. E., K. A. Emanuel, and C. M. Snyder, 2001: Idealized adaptive observation strategies for improving numerical weather prediction. *J. Atmos. Sci.*, **58**, 210–232, https://doi.org/10.1175/1520-0469(2001)058<0210:IAOSFI>2.0.CO;2.

Munoz-Sabater, J., H. Lawrence, C. Albergel, P. de Rosnay, L. Isaksen, S. Mecklenburg, Y. Kerr, and M. Drusch, 2019: Assimilation of SMOS brightness temperatures in the ECMWF IFS. ECMWF Tech. Memo. 843, 38 pp., http://doi.org/10.21957/qq4v2o7oy.

Orth, R., E. Dutra, I. F. Trigo, and G. Balsamo, 2017: Advancing land surface model development with satellite-based Earth observations. *Hydrol. Earth Syst. Sci.*, **21**, 2483–2495, https://doi.org/10.5194/hess-21-2483-2017.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quart. J. Roy. Meteor. Soc.*, **127**, 279–304, https://doi.org/10.1002/qj.49712757202.

——, 2006: Predictability of weather and climate: From theory to practice. *Predictability of Weather and Climate*, T. N. Palmer and R. Hagedorn, Eds., Cambridge Press, 1–29.

——, 2019a: The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quart. J. Roy. Meteor. Soc.*, **145**, 12–24, https://doi.org/10.1002/qj.3383.

——, 2019b: Stochastic weather and climate models. *Nat. Rev. Phys.*, **1**, 463–471, https://doi.org/10.1038/s42254-019-0062-2.

Rasmussen, C. E., and C. K. I. Williams, 2006: *Gaussian Processes in Machine Learning*. MIT Press, 248 pp.

Richardson, D. S., 2000: Skill and economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Rodwell, M. J., D. S. Richardson, D. B. Parsons, and H. Wernli, 2018: Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bull. Amer. Meteor. Soc.*, **99**, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1.

Saetra, Ø., H. Hersbach, J. R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.*, **132**, 1487–1501, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2.

Sutton, C., T. M. Hamill, and T. T. Warner, 2006: Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Wea. Rev.*, **134**, 3174–3189, https://doi.org/10.1175/MWR3248.1.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bull. Amer. Meteor. Soc.*, **97**, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1.

Tennant, W., and S. Beare, 2014: New schemes to perturb sea-surface temperature and soil moisture content in MOGREPS. *Quart. J. Roy. Meteor. Soc.*, **140**, 1150–1160, https://doi.org/10.1002/qj.2202.

Warner, T. T., Ed., 2011: Ensemble methods. *Numerical Weather and Climate Prediction*, Cambridge Press, 252–280.

Weijs, S. V., and N. van de Giesen, 2011: Accounting for observational uncertainty in forecast verification: An information- theoretical view on forecasts, observations, and truth. *Mon. Wea. Rev.*, **139**, 2156–2162, https://doi.org/10.1175/2011MWR3573.1.

Weisheimer, A., S. Corti, T. Palmer, and F. Vitart, 2014: Addressing model error through atmospheric stochastic physical parametrizations: Impact on the coupled ECMWF seasonal forecasting system. *Philos. Trans. Roy. Soc.*, **372**, 20130290, http://doi.org/10.1098/rsta.2013.0290.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

——, 2016: "The stippling shows statistically significant grid points": How research results are routinely overstated and overinterpreted, and what to do about it. *Bull. Amer. Meteor. Soc.*, **97**, 2263–2273, https://doi.org/10.1175/BAMS-D-15-00267.1.

Yamaguchi, M., S. T. Lang, M. Leutbecher, M. J. Rodwell, G. Radnoti, and N. Bormann, 2016: Observation-based evaluation of ensemble reliability. *Quart. J. Roy. Meteor. Soc.*, **142**, 506–514, https://doi.org/10.1002/qj.2675.

Zhang, C., 2013: Madden–Julian oscillation: Bridging weather and climate. *Bull. Amer. Meteor. Soc.*, **94**, 1849–1870, https://doi.org/10.1175/BAMS-D-12-00026.1.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–84, https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.