2011

þÿ T h e U . S . A i r F o r c e W e a t h e r A g e n c y s m ensemble: scientific description and performance results

Hacker, J.P.

þÿ T e l l u s (2 0 1 1) , 6 3 A , p p . 6 2 5 6 4 1
http://hdl.handle.net/10945/47185

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

Tellus (2011), 63A, 625–641
Printed in Singapore. All rights reserved

TELLUS

The U.S. Air Force Weather Agency's mesoscale ensemble: scientific description and performance results

By J. P. HACKER[1]*, S.-Y. HA[2], C. SNYDER[2], J. BERNER[2], F. A. ECKEL[3], E. KUCHERA[4], M. POCERNICH[2], S. RUGG[4], J. SCHRAMM[2] and X. WANG[5], [1]*Naval Postgraduate School, Monterey, CA, USA;* [2]*National Center for Atmospheric Research, Boulder, CO, USA;* [3]*National Weather Service Office of Science and Technology, Silver Spring, MD, USA;* [4]*Air Force Weather Agency, Bellevue, NE, USA;* [5]*University of Oklahoma, Norman, OK, USA*

(Manuscript received 14 April 2010; in final form 1 December 2010)

ABSTRACT

This work evaluates several techniques to account for mesoscale initial-condition (IC) and model uncertainty in a short-range ensemble prediction system based on the Weather Research and Forecast (WRF) model. A scientific description and verification of several candidate methods for implementation in the U.S. Air Force Weather Agency mesoscale ensemble is presented. Model perturbation methods tested include multiple parametrization suites, land-surface property perturbations, perturbations to parameters within physics schemes and stochastic 'backscatter' stream-function perturbations. IC perturbations considered include perturbed observations in 10 independent WRF-3DVar cycles and the ensemble-transform Kalman filter (ETKF). A hybrid of ETKF (for IC perturbations) and WRF-3DVar (to update the ensemble mean) is also tested. Results show that all of the model and IC perturbation methods examined are more skilful than direct dynamical downscaling of the global ensemble. IC perturbations are most helpful during the first 12 h of the forecasts. Physical parametrization diversity appears critical for boundary-layer forecasts. In an effort to reduce system complexity by reducing the number of suites of physical parametrizations, a smaller set of parametrization suites was combined with perturbed parameters and stochastic backscatter, resulting in the most skilful and statistically consistent ensemble predictions.

1. Introduction

Short-range, mesoscale ensemble prediction has been a topic of applied research for well over a decade, but major fundamental advances have been elusive. Questions regarding the relative importance of initial-condition (IC) and model errors, and how they relate to temporal and spatial scales, still remain. Methods to account for mesoscale sources of uncertainty might still be considered immature.

The U.S. Air Force (USAF) Weather Agency (AFWA) has been pursuing numerical weather prediction ensemble technology, often jointly with the assistance of the U.S. Navy's Fleet Numerical Meteorology and Oceanography Center, since early 2007. Through this endeavour, AFWA has leveraged research from the National Center for Atmospheric Research (NCAR), accomplishments from the North American Ensemble Forecast System (NAEFS; http://www.emc.ncep.noaa.gov/gmb/ens/NAEFS.html), among other national academic institutions and forecast operations. As of this writing, AFWA is on the brink of operationally implementing an ensemble of international global ensembles from centres located in the United States and Canada. In parallel, AFWA and NCAR have been developing a mesoscale ensemble based on the limited-area Weather Research and Forecast (WRF) model. High-impact aviation weather parameters and surface/near-surface weather phenomena have been the initial focus of this effort.

The mesoscale ensemble was evaluated throughout 2009 by forecasters and meteorologists of the U.S. armed forces. Sufficient skill and reliability was attained, warranting formal operational implementation on domains worldwide by late 2010. Indeed, the operational weather squadron predicting for the Middle East, northern Africa and western Asia already uses prototypical dust-lofting probabilistic forecasts in their decision process, both to focus their forecast efforts and also to confirm regional deterministic prognoses.

*Corresponding author.
e-mail: jphacker@nps.edu
DOI: 10.1111/j.1600-0870.2010.00497.x

Compared to other operational centres, AFWA has a relatively small user base with more narrowly defined needs. Development of the mesoscale ensemble system has therefore concentrated on the following areas relevant to aviation support: (1) predictions for primarily the lower atmosphere (winds) with a broadening scope to the mid-troposphere (e.g. visibility); (2) multiple time scales (0–60 h) and fine spatial scales (ultimately <5 km but in this work we test to 15 km and verify only 45 km); (3) the ability to run decision aids from numerical ensemble forecasts and (4) products to complement global ensemble predictions.

The immediate goal of this work is to find a combination of IC perturbation and model 'perturbations' that produces an effective ensemble relevant to USAF aviation needs. More established approaches using direct dynamical downscaling from a global ensemble, and varying physics suites within the WRF model (multiphysics ensemble akin to those investigated by Stensrud et al., 2000; Ziehmann, 2000; Hou et al., 2001; Grimit and Mass, 2002; Stensrud and Yussouf, 2003; Eckel and Mass, 2005; Clark et al., 2008), are useful baselines for mesoscale ensemble prediction.

While multiphysics ensembles have been successful, the USAF employs many secondary models, or decision aids, and each of these secondary models must be calibrated for each suite of physical parametrizations used in the ensemble. Such tuning would require substantial resources, thus motivating approaches to account for model uncertainty that involve either a single parametrization suite or a greatly reduced set of parametrization suites. Compared to a baseline of 10 parametrization suites, here we consider perturbing parameters within a single parametrization suite, and perturbing parameters within three parametrization suites. We also examine stochastic stream-function perturbations. An over-arching goal of AFWA development efforts is to find the least complex system delivering useful skill based on metrics that are relevant to USAF operations.

With goals of reducing complexity and effectively accounting for both IC and model uncertainty in a mesoscale ensemble prediction system, we report results from several ensemble prediction experiments. We verify lower-tropospheric ensemble predictions from several experiments, looking for significant differences resulting from different techniques for ensemble prediction. Direct dynamical downscaling and a multiphysics ensemble provide baselines for skill. Similar to Bowler et al. (2008) (the Met Office Global and Regional Ensemble Prediction System) and Wei et al. (2008) (the NCEP Global Ensemble Forecast System) we report the ensemble performance to help other developers of short-range, mesoscale ensemble prediction systems make informed decisions about ensemble design. The practical goal, of providing context for more scientific examination of ensemble methods for the WRF, is met with this overview paper which is necessarily brief in details on any single method. In Section 2 a brief description of each method, and references for additional information, are given.

The next section contains a description of the experiments and the ensemble construction. Section 3 contains descriptions of evaluation methods and data sets. Sections 4 and 5 present comparisons between different groups of ensembles; Section 4 compares forecasts using various methods to account for mesoscale IC uncertainty or model uncertainty and Section 5 explores methods aimed at reducing the number of physics variations in an ensemble. Results are summarized in Section 6.

2. Experiment description

Objective and general methods for designing an ensemble to meet even a well-defined and specific need are non-existent. Instead we follow typical practices and adopt an approach of a priori reasoning, computational experimentation and a posteriori empiricism to find the ensemble that best meets USAF needs. To address mesoscale model, lower boundary and IC uncertainty, experiments tested a wide range of perturbations.

Computational resources limit the ensemble to 10 members in the AFWA implementation. Development and testing is consequently on 10-member ensembles. Larger ensembles were also tested for specific experiments but we do not present those results. Although experimentation uses both a 45-/15-km grid spacing one-way nested domain configuration and a single 45-km domain, results presented here are from only the 45-km domain; we focus on the methods rather than the effects of resolution or ensemble size. Forecasts are initialized twice daily, at 0000 and 1200 UTC, to make 60-h predictions. Forecasting on alternate days extends the experiments to a greater variety of weather scenarios while keeping computational demands manageable. As described later, we experimented on two domains: a continental U.S. (CONUS) and an East-Asian domain. Both are shown in Fig. 1, with locations of upper-air sonde observations.

All ensembles use NCEP's Global Ensemble Forecast System (GEFS; Wei et al., 2008) for lateral boundary conditions (LBCs). GEFS is constructed from the Global Forecast System (GFS) model and an ensemble transform (ET) technique (Bishop, 1999). The ET implementation in GEFS includes a regional initial perturbation scaling to account for regional differences in analysis error variance from the operational three-dimensional-var scheme.

Below we briefly describe the baseline ensemble and each perturbation method as implemented in these experiments. For reference, characteristics of each ensemble are summarized in Table 1, which shows the relationships between the different ensembles and perturbation methods. Because *Cntl* uses the most straightforward formulation, all of the other ensembles can be interpreted in the context of adding skill to it.

2.1. Baseline downscaled ensemble

Perhaps the simplest approach to mesoscale ensemble prediction is through direct dynamical downscaling of a global ensemble

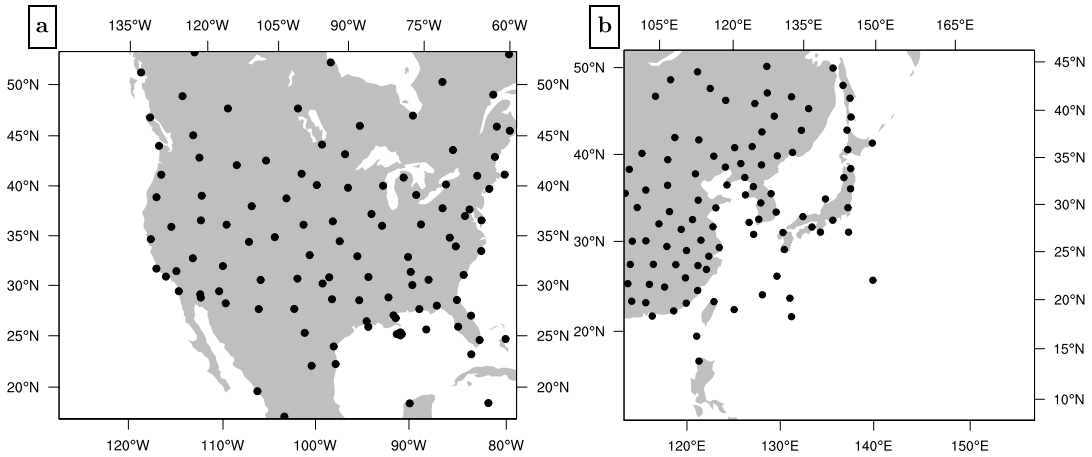

Fig. 1. The continental U.S. (CONUS) domain in (a) and the East-Asian domain in (b). Both use $\Delta X = 45$ km. Dots show upper-air sonde locations.

Table 1. Experiment list with names and symbols (colour and black) used in later figures.

Symbol	Symbol	Name	Description	Physics (suite no.)
○	○	*Cntl*	Single suite of control physics	Single control (6)
△	△	*Phys*	Ten physics suites	10 suites (1–10)
□	□	*PO*	Perturbed observations, independent WRF-3DVar	10 suites (1–10)
●	●	*ETKF*	Ensemble-transform Kalman Filter	10 suites (1–10)
▲	▲	*Hybrid*	WRF-3DVar (mean) and ETKF (perturbations)	10 suites (1–10)
■	■	*Stoch*	Stochastic stream-function perturbations (‘back-scatter”)	Single control (6)
▽	▽	*Param*	Perturbed physics parameters	Single control (6)
▼	▼	*LMP2*	Limited (three) physics suites; perturbed physics parameters	3 suites ($6,7^*_{○}9$)
✳	✳	*LMP2_Stoch*	*LMP2* plus *Stoch*	3 suites ($6,7^*_{○}9$)

Notes: All experiments use the GEFS for lateral boundary conditions, and also land-use perturbations. The Physics column notes in parentheses the specific suites from Table 2. The asterisk on Member 7 in ensembles *LMP2* and *LMP2_Stoch* denotes that the WSM5 microphysics scheme in Member 7 was replaced with the Thompson scheme.

by nesting a limited-area model within each global member. A first-order requirement of a regional ensemble is that it performs at least as well, subject to selected norms, as a direct downscaling with a well-tested and accepted implementation of the limited-area model. One could further argue that any regional ensemble should only be considered if it outperforms a global, usually coarser-resolution, ensemble under the same metrics. We avoid this argument and instead assume that regional, mesoscale, ensembles can be of value to forecasters and decision makers simply because they provide interpretable mesoscale realism to a forecaster.

A downscaled global ensemble (denoted *Cntl* because it uses the control/operational suite of physics) with the Advanced Research version of the WRF version 3.1 limited-area model (Skamarock et al., 2008) is the point of comparison for all other experiments. Each member uses the AFWA operational suite of physics (see next section), and perturbations to land-use tables that affect surface drag and energy balances, as described below. All of the other methods implemented and tested will be evaluated relative to *Cntl*.

The GEFS contained 21 members during the experiment period reported here, centred in phase space via a simplex method (cf. Wang et al., 2004; Wei et al., 2008). *Cntl* makes use of the first 10 members of GEFS; its IC mean is approximately the mean of the full 21-member ensemble, with differences attributable to sampling error. Each mesoscale member is consistently associated with the same GEFS member, both in initial conditions and for LBCs.

The *Cntl* ensemble also uses perturbations to properties of the lower boundary. Eckel and Mass (2005) describe a technique to assign perturbations to land-surface parameters. Albedo, soil moisture availability and roughness length are perturbed with random draws from Γ-like distributions, with distribution parameters chosen through physical arguments and empirical data. Separate land-surface tables are generated for each ensemble member, and do not change throughout the experiment. Applying land-use perturbations to 28 forecasts during October 2006 over East Asia led to slight improvements. Figure 2 shows that error measured by the root-mean-square ensemble-mean error (RMSE), where an ensemble-mean error is an observation

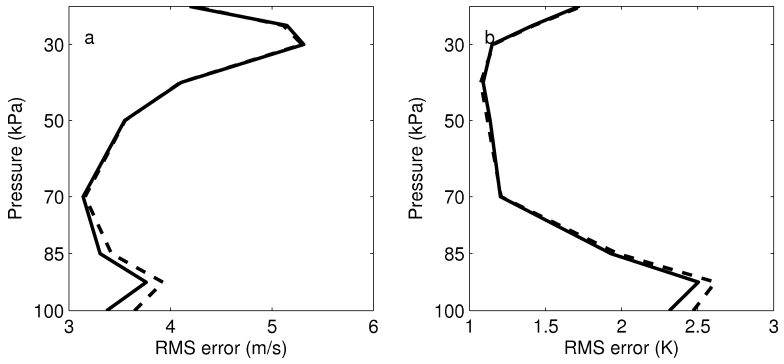

Fig. 2. Root-mean-square ensemble-mean error (RMSE) of the control physics and land-use perturbation ensemble (*Cntl*; solid) and an ensemble without land-use perturbations (dashed). Shown are (a) zonal wind component and (b) temperature for 28 forecasts at 48-h lead time during October 2006 over East Asia.

minus the ensemble-mean forecast [$o - \overline{f}$; see eq. (1) in Section 3.1 for clarification] is reduced. Error reductions are small and confined to the lower troposphere. Although Fig. 2 shows results for the zonal wind component and temperature at a 48-h lead time, results are qualitatively similar for meridional wind and water vapour mixing ratio, and for other forecast lead times. Overall, as we will see, the effects of land-use perturbations are small compared to the effects of varying land-surface models.

2.2. Representing mesoscale IC uncertainty

The initial conditions in *Cntl* are inherited from the NCEP global ensemble (GEFS). A potential improvement to *Cntl* is to employ ICs that are generated under the the dynamics of the WRF itself and which account, at least approximately, for both the details of the observation network and the temporal variations in analysis uncertainty. We test three approaches for producing an ensemble of ICs in the mesoscale model.

The first approach is the Monte-Carlo methodology of Houtekamer and Derome (1995). They proposed performing parallel data-assimilation cycles for each member using simulated observations derived by perturbing the real observations with random noise consistent with the observation-error statistics assumed in the data-assimilation system. Hamill et al. (2000) showed in a simple model that this perturbed-observation method (*PO* hereafter) produced probabilistic forecasts superior to those from either bred vectors or approximate singular vectors. The *PO* method is designed to produce, at least approximately, a random sample from the distribution of atmospheric states given the latest observations. Houtekamer and Derome (1995) and Hamill et al. (2000) provide further discussion, while Burgers et al. (1998) show that *PO* does in fact sample from the correct distribution in the case that observation errors are Gaussian, the system is linear and the ensemble size is large. Because each member in *PO* performs data assimilation independently, the ICs will differ from those of *Cntl* in both the mean and deviations about the mean. A three-dimensional variational data assimilation system is used here (WRF 3DVar; Skamarock et al., 2008).

The second method is the ensemble-transform Kalman filter (*ETKF*) (Wang and Bishop, 2003; Wang et al., 2004). In the *ETKF*, the initial perturbations are generated by updating the forecast perturbations with a transformation matrix. The transformation depends on the observation operator and error variances, and is chosen to yield an updated ensemble with covariance approximately equal to the analysis error covariance that would result from a Kalman filter, given the covariance derived from the forecast ensemble. In our study, the *ETKF* update is performed domain-wide without covariance localization (Houtekamer and Mitchell, 2001; Hamill et al., 2001), and therefore systematically underestimates the analysis-error variance owing to the small ensemble size. We implement an adaptive inflation factor as in Wang and Bishop (2003) to ameliorate the systematic analysis-error variance underestimation. Wang and Bishop (2003) demonstrated the efficacy of the ETKF with inflation in a global forecast context. The inflation factor is averaged over the most-recent four cycles (48 h total) to smooth it. Because of the small ensemble and lack of covariance localization, *ETKF* here cannot be used effectively for data assimilation, thus in this implementation the *ETKF* initial perturbations are added to the GEFS ensemble mean at each analysis time. Because no data assimilation is performed, the *ETKF* is computationally less expensive than *PO*.

A third method is to update the ensemble mean using a hybrid ensemble-variational assimilation scheme for the WRF, and compute the ensemble perturbations using the ETKF. This approach, taken by Wang et al. (2008a,b), leads to the *Hybrid* ensemble. Assimilation employs extended control variables in the variational update (Lorenc, 2003). This allows the ensemble covariance to be combined, in a weighted sum, with the stationary and isotropic background error covariance estimate derived from differences between 24- and 12-h WRF forecasts (Parrish and Derber, 1992) using the control physics. These flow-dependent covariances are expected to improve analyses, and hopefully forecasts, compared to 3DVar using static and isotropic covariances. Here we equally weight the two covariance estimates. Wang et al. (2008a) give further details, and Wang et al. (2007) showed that the ensemble covariances are effectively localized in space.

Table 2. Configuration of multiphysics ensemble.

Member	Land Surface	PBL	Microphysics	Cumulus	Long-wave	Short-wave
1	Thermal	YSU	Kessler	KF	RRTM	Dudhia
2	RUC	MYJ	Eta	KF	RRTM	Dudhia
3	Thermal	MYJ	WSM6	KF	RRTM	CAM
4	Noah	MYJ	Kessler	BM	CAM	Dudhia
5	Noah	MYJ	Lin	Grell	CAM	CAM
6	Noah	YSU	WSM5	KF	RRTM	Dudhia
7	Noah	MYJ	WSM5	Grell	RRTM	Dudhia
8	RUC	YSU	Lin	BM	CAM	Dudhia
9	RUC	YSU	Eta	BM	RRTM	CAM
10	RUC	MYJ	Thompson	Grell	CAM	CAM

Note: Member 6 uses the same physics suite as the operational configuration at AFWA.

The IC-perturbation methods studied here use multiple physics suites, and their skill can thus be evaluated relative to the multiphysics ensemble described in the next section. The *PO*, *ETKF* and *Hybrid* use observations at 0000 and 1200 UTC to update first-guess perturbations taken from the 12-h ensemble forecasts, resulting in new perturbations every 12 h. Balloon-borne soundings, cloud-drift winds from GOES water vapour imagery (Velden et al., 1997) and Aircraft Communications Addressing and Reporting System (ACARS; Lord et al., 1984) in situ reports were used in all experiments. Experiments over CONUS used surface observations, but the East Asian experiments did not. Observation error-variance values, where needed, were borrowed from NCEP estimates. These are discussed in more detail later.

2.3. Simulating model uncertainty

An ensemble that attempts to account for model uncertainty can be easily created by choosing distinct physics suites for each ensemble member. Physics variations may include subgrid scale closure (PBL, microphysics and deep convection), forcing (radiative transfer) and lower boundary conditions (land-surface model or a relaxation scheme). Eckel and Mass (2005) argue that this is one way to generate models with different attractors, which may be beneficial because no single model reproduces the atmosphere's attractor. Their results show that multiphysics ensembles contribute important diversity to an ensemble, but that including entirely different modelling systems (such as the WRF and MM5) in an ensemble leads to still more useful information. Several other studies, for example, Stensrud et al. (2000); Ziehmann (2000); Hou et al. (2001); Grimit and Mass (2002); Stensrud and Yussouf (2003); Eckel and Mass (2005); Clark et al. (2008), have demonstrated the utility of model diversity in ensembles.

Table 2 summarizes the parametrization suites for ensemble *Phys*. By selecting schemes that fundamentally differ from each other, we made an heuristic attempt to include as much diversity in classes of physics schemes as possible. Considering only the physics schemes supported with the release of WRF version 3.1, the number of possible combinations of schemes is (4 land surface)$\times$(4 PBL)$\times$(7 microphysics)$\times$(4 cumulus)$\times$(3 long-wave)$\times$(3 short-wave) $=$ 4032. Because physics schemes are in practice tuned as a suite, many combinations do not work well or are difficult to use together. We found that the suites in Table 2 run stably and produce reasonable forecasts. Details and references for all the physics are in Skamarock et al. (2008).

Imposing perturbations to parameters within a single set of physics schemes produces an alternative denoted *Param*. Murphy et al. (2004) and Stainforth et al. (2005) found climate-prediction sensitivity to parameter perturbations, and also found that model quality could degrade with some parameter choices. It is not clear whether model error in the faster time scales characterizing NWP can be simulated by varying parameters. Bowler et al. (2008) found a small positive impact when applying an auto-regressive process to parameters in the Met Office ensemble prediction system, but that using multiple physics schemes led to greater benefit. In this issue, Hacker et al. (2011) describes our approach more thoroughly. Briefly, single parameters are chosen in each of the PBL, microphysics, cumulus and short-wave radiation schemes in Member 6 from Table 2. Member 6 uses the AFWA operational physics suite, which is the same used in ensemble *Cntl*. Parameter choices are based on known sensitivity as reported in the literature, and subsequent sensitivity tests. Ten parameter sets, each corresponding to a unique ensemble member, are chosen with a space-filling Latin Hypercube Sampling (Santer and Williams, 2003). In this paper we simply compare the predictions from *Param* with others.

The stochastic kinetic energy backscatter[1] scheme takes yet another approach, aiming to represent model uncertainty resulting from interactions with unresolved scales. It is based on the notion that the turbulent dissipation rate is the difference

[1] The approach taken here is not formally backscatter because it lacks an explicit link between dissipation and perturbations, but we retain this terminology for consistency with published literature.

between upscale and downscale spectral transfer, with the upscale component being available to the resolved flow as a kinetic energy source (Shutts, 2005). To simulate a stochastic kinetic energy source, we follow Berner et al. (2009) and introduce random stream-function and temperature perturbations with a prescribed kinetic energy spectrum. This approximate backscatter was shown by Shutts (2005) and Berner et al. (2009) to be just as effective as dissipation-dependent backscatter. The power-law exponent was estimated from coarse-grained high-resolution model output. Spatial correlations in the random pattern are generated by expanding the stream-function forcing in spectral space and evolving each wavenumber as a first-order auto-regressive process. This allows full control over the spatial and temporal characteristics of the perturbations, and in practice the ensemble spread from the perturbations can be tuned. The stochastic kinetic energy backscatter scheme, assuming spatially constant dissipation rate as assumed here, has been shown to improve the skill in the ECMWF ensemble forecasting system (Berner et al., 2009). Its implementation and performance in the AFWA mesoscale ensemble are discussed in detail in Berner et al. (2011).

A reduced set of three physics suites, combined with parameter perturbations, was also tested to explore the potential for less complexity. Clear theoretical guidance for choosing members from the complete list in Table 2 is lacking. Under the constraint that AFWA's operational configuration (Member 6) be included, we initially chose configurations 3 and 9 to complete the ensemble based on the following objective goals: (1) Exclude members that have especially large deterministic errors, as measured by the RMS and mean of observation minus forecast ($o - f$) values, averaged over the domain and over each experiment period. (2) Select members whose differences have as little correlation as possible, as measured by the temporal correlation between paired $o - f$ time series at a given lead time. Member correlations with other members were summarized by averaging squared correlations over all observations and then summing over all other members. (3) Select a subset of members whose variance is as large as possible, by computing the variance for all three-member subensembles that include the operational configuration.

The thermal (multilayer force-restore) land-surface scheme leads to large near-surface summer-time biases (not shown) in Member 3. We thus chose Member 7, which differs only slightly from Member 3 based on criteria (2) and (3) above, to replace it. Switching the WSM5 for the Thompson microphysics scheme introduced further variability.

Hacker et al. (2011), in this issue, details the choice of parameters for Member 6. Parameters to perturb the six additional physics schemes for Members 7 and 9 were also chosen based on literature reviews and sensitivity studies, and are presented in the Appendix. Three values for each parameter are used here: the default, a high value and a low value. A coin flip determined which perturbed member adopted the high or low value of each parametrization. This ensemble is termed Limited Multi-Physics, Multi-Parameter (*LMP2*).

Finally, the techniques in *Stoch* and *LMP2* were combined to create the ensemble *LMP2_Stoch*. We show below that this combination produces the most skilful probabilistic forecasts of any tested here.

3. Evaluation methods

3.1. Observations and metrics

Ensemble performance evaluation follows typical probabilistic verification practices, and includes metrics to assess statistical consistency, reliability and resolution. Rank histograms and reliability diagrams separate reliability from resolution; Brier scores and continuous rank-probability scores (CRPS) summarize the joint contribution of reliability and resolution. For a detailed discussion of these metrics, we refer the reader to Jolliffe and Stephenson (2003).

Another measure of reliability is the degree of consistency between ensemble spread and error. A reliable ensemble will exhibit approximate agreement between root-mean-square ensemble-mean error (RMSE) and 'total spread', which includes both ensemble spread and observation error. This approximate agreement expresses the degree to which the ensemble can on-average predict the observation distribution, and can be expressed as

$$\left[\frac{1}{N-1}\sum_{n=1}^{N}\left(o_n-\overline{f_n}\right)^2\right]^{1/2}\approx\left[\frac{1}{N-1}\sum_{n=1}^{N}\left(\sigma_{f,n}^2+\sigma_{o,n}^2\right)\right]^{1/2}, \tag{1}$$

where RMSE of the ensemble mean is the left-hand side, total spread is the right-hand side, the subscript $n = 1, \ldots, N$ indexes the total number of verifying observations for the experiment valid at a particular forecast lead time, $\overline{f_n}$ is the ensemble-mean forecast, $\sigma_{f,n}^2$ is ensemble variance and $\sigma_{o,n}^2$ is observation-error variance. Here we will evaluate relative consistency between the different ensembles.

In a multiphysics or perturbed-parameter ensemble, each member can be differently biased. Defining bias to be the experiment-mean error as a function of observing station, pressure level (or surface) and forecast lead time, we remove the bias of each forecast before computing scores or spreads. That is, we use corrected individual-member forecasts at a given lead time $f_k' = f_k - \overline{(o - f)}$, where k indexes observations or forecasts at particular horizontal location, level and ensemble member, and the average is over available instances of verifying observations.

Observation error estimates can also be considered in the verification. Estimating observation error values is generally difficult, but within a data assimilation context it is possible to obtain values consistent with a particular model (e.g. Desroziers et al., 2005). We test with values of σ_o estimated at NCEP, with the

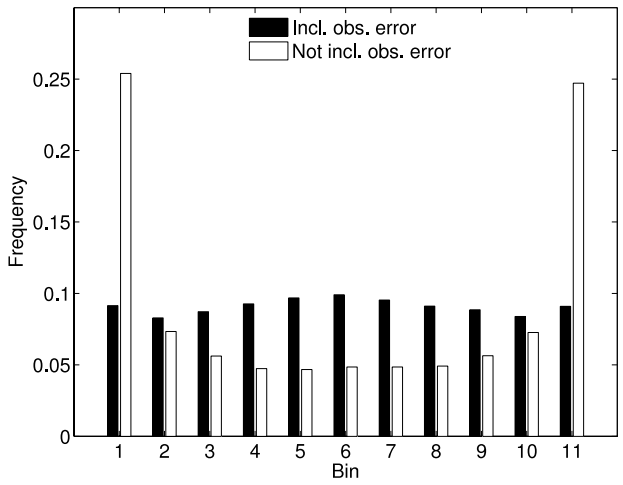

Fig. 3. Rank histograms from 24-h 10-m wind speed predictions from the multiphysics (*Phys*) ensemble, including (black) and not including (white) observation errors. The 10-m wind speed observation error standard deviation is approx. 1.6 m s^{-1}.

understanding that they may not be accurate for our system. During the verification we draw a simulated, random observation-error value $\sim \mathcal{N}(0, \sigma_o)$ (a sample mean of zero is explicitly enforced) and add it to each ensemble member forecast f before computing an error $(o - f)$. Then scores are computed as usual.

Including observation error can change the conclusions one draws from verification. A comparison of rank histograms for 24-h 10-m wind speed predictions from ensemble *Phys*, with and without estimated observation errors with $\sigma_o \approx 1.6$ m s^{-1}, is shown in Fig. 3. Without observation error, the observation is outside of the extreme ensemble values in approximately half of the verification sample, and we conclude that the ensemble is underdispersive or has conditional biases. Adding random observation errors to the predictions produces a rank histogram that is much flatter and appears reasonably reliable although slightly overdispersive. However given that we cannot know the observation error precisely, and the results are sensitive to it, conclusions about reliability (and performance in general) cannot be made with certainty.

Although consideration of observation errors is desirable, we lack rigorous estimates of observation errors. Values of σ_o obtained from NCEP range from 1.1 to 3.3 m s^{-1} for wind components, and 1–2 K for temperatures observed with balloon-borne sondes (Fig. 1) and at surface observing stations; these values are similar in magnitude to RMSE and spread (see for example error plots in Fig. 2), suggesting that NCEP σ_o values are too high for the present system. We therefore choose to ignore the effects of observation errors on the verification hereafter, and focus on the relative skill among the different ensembles.

The following steps were taken to create intervals describing uncertainty in the skill scores, and the differences between scores for different ensembles (for further details refer to Wilks, 2006). The score or difference was repeatedly calculated on data randomly sampled from the original data. The sampling was done with replacement and for each sample, the same set of dates was used from each experiment. The intervals depict the 5th and 95th percentiles of the scores and differences. Variability of a set of predictions over many weather scenarios is much greater than variability of the differences between two models or ensembles over the same weather scenarios. Uncertainty estimates on the scores of two different ensembles hides this variability of the ensemble differences and will lead to the conclusion that two ensembles cannot be judged as different. It is more appropriate to estimate uncertainty from the distributions of score differences instead, therefore accounting for the internal variability in the predictions. We take the latter approach.

3.2. Data and evaluation periods

Evaluation is restricted to the balloon-borne upper-air soundings and surface observing stations. The primary test period over CONUS is during November–December 2008. The domain contains approximately 3000 surface observing stations and 100 upper-air stations. October 2006 over East Asia provides data for evaluation in a different regime. Precipitation resulted primarily from warm cloud microphysics, but frozen hydrometeors were certainly present; Typhoon Soulik tracked through the domain during 14–16 October. For East Asia, approximately 500 surface observing stations and 30 upper-air stations were available. Two test periods in different synoptic regimes give the same performance rank for each ensemble, but the absolute values of the scores are different. Below we show results from the CONUS case for most of the ensembles, where there are more observations and greater regime variability, and therefore more statistical significance. To compare the ETKF and the hybrid 3DVar/ETKF we use the East Asian case. In all instances we verify with observations valid at 0000 and 1200 UTC. In each instance where skill comparisons are made, the verification data set was determined by the intersection of forecasts dates available. The comparisons can be either one or two months of alternate-day forecasts (32 or 62 forecasts, respectively for CONUS and 28 for East-Asia). The specific periods are noted in figure captions.

4. Results

In this section we demonstrate that including either mesoscale IC uncertainty or simulated model uncertainty can improve forecasts relative to *Cntl*. Further, most of the techniques lead to forecasts that outperform *Phys*. Combining a limited number of physics variations with multiple parameters and stochastic physics appears to give superior performance. Differences between near-surface and 70-kPa skill shows that a single mesoscale ensemble technique does not sufficiently capture uncertainty for all forecast parameters.

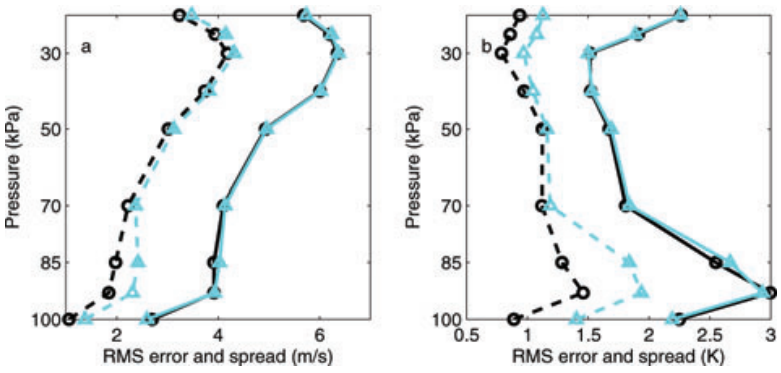

Fig. 4. Root-mean-square ensemble-mean error (RMSE; solid curves) and total spread (dashed curves) of *Cntl* (circle) and *Phys* (triangle). Shown are (a) zonal wind component and (b) temperature for 62 forecasts at 48-h lead time during November 2008–January 2009 over the continental United States.

4.1. Comparison to Cntl

Use of multiple physics suites improves several aspects of the forecast, and shows particular benefit in the PBL. RMSE and total spread for *Cntl* and *Phys* shows that statistical consistency is superior in *Phys* (Fig. 4). *Phys* shows greater spread than *Cntl* for both wind (panel a) and temperature (panel b), with the most notable differences in the PBL.

Rank histograms show that *Phys*, *PO* and *Param* all improve reliability of 2-m temperature and 10-m wind-speed predictions compared to *Cntl* (Fig. 5). Relative performance among the different ensembles is the same for both temperature (panels a and b) and wind speed (panels c and d). All ensembles show endemic underdispersion when observation error is not considered (see Section 3.1), but become slightly more reliable as the forecast lead time increases from 12 h (panels a and c) to 48 h (panels b and d). Similar improvement in reliability among all the ensembles suggests that the LBCs, which determine large-scale growth in ensemble spread, are primarily responsible. Differences between the ensembles are smaller. Ensemble *Param* provides only slight improvements over *Cntl*. Ensembles *Phys* and *PO* offer further reliability from the physics diversity, and *PO* shows the short lead-time benefit of mesoscale IC variability from the perturbed observation approach (Fig. 5a).

The CRPS is a generalization of the Brier score to all thresholds in the observed distribution, and includes contributions from both reliability and resolution. We verified (not shown) that Brier scores for predictions exceeding individual thresholds ranging from the 25th to the 75th observation-distribution percentile give the same skill ranking among the ensembles, and the CRPS can be confidently interpreted. We are interested in the difference between a particular ensemble scheme and the straightforward *Cntl*. The CRPS is negatively oriented and for presentation we reverse the difference so that an improvement over *Cntl* is shown as a positive value.

CRPS results show that the greatest benefit of multiple physics is realized at the surface, but that multiparameter techniques can be competitive aloft (Fig. 6). *Phys* (triangle) and *PO* (square) offer similar improvements over *Cntl* for 2-m temperature (panel a) and 10-m wind speed (panel c). At initialization near the surface, *PO* shows additional benefit from introducing explicit mesoscale IC perturbations. Multiple PBL schemes and land-surface models can introduce diversity within the ensemble at fast time scales near the surface. Skill relative to *Cntl* diminishes with forecast lead time, either because larger scale uncertainty becomes more important or diversity in the PBL and soil states of *Cntl* (and *Param*) has grown.

At 70 kPa (Figs. 6b and d) CRPS differences from *Cntl* are smaller and uncertainty in those differences, shown by the vertical lines, is greater. Few differences can be accepted as meaningful. *Phys* and *PO* both show slight deterioration in 70-kPa temperature CRPS (panel b) relative to *Cntl* and *PO* shows skill reductions in 70-kPa wind speed (panel d) during the first 36-h lead time. At 70 kPa *Param* (inverted triangle) shows no skill deterioration, but the longer vertical lines shows that the distribution of skill differences is wide.

Ensemble *PO* shows the benefit of data assimilation at very short time scales, but it is not perfect. A perfect fit of all ensemble members to observations would give a perfect CRPS. Given an optimal data assimilation system and a perfect model, observation errors prevent a perfect fit and a perfect CRPS. In practice, suboptimalities and model deficiencies limit the CRPS further. Ensemble *PO* still shows much of the benefit of ensemble data assimilation at initialization. The benefit is quickly lost, and it is unclear why wind speed at 70 kPa does not show improved CRPS. This topic would require further investigation.

Improved reliability from using multiple physics suites accounts for some of the improvements in CRPS at the surface. Reliability for near-surface predictions of exceeding the 75th percentile of observation distributions at each individual observing location show (Fig. 7) that ensembles *Phys* (triangle), *PO* (square) and *Param* (inverted triangle) all offer greater reliability than *Cntl* (circle). *Phys* and *PO* show greatest reliability for 2-m temperature at both 12 h (panel a) and 48 h (panel b) lead times. They are most notably more skilful at predicting threshold exceedance with high probability, indicating skill in the highest temperature quartile. At lower probability the ensembles perform similarly, indicating similar skill when exceedances are predicted with low probability. Most forecasts are for low

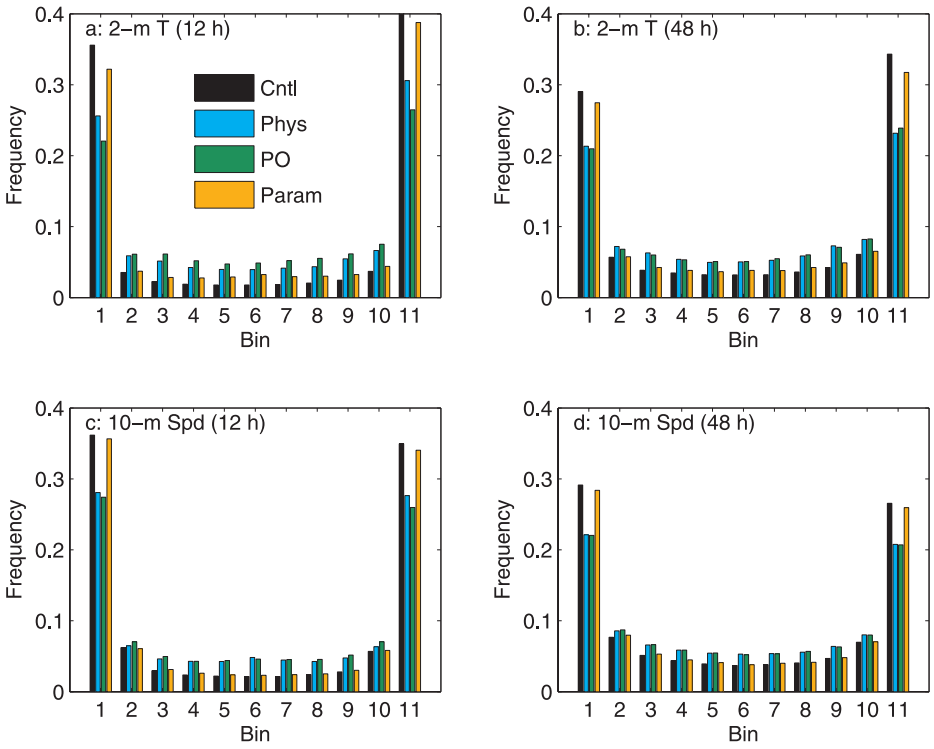

Fig. 5. Rank histograms from (a) 12-h 2-m temperature, (b) 48-h 2-m temperature, (c) 12-h 10-m wind speed and (d) 48-h 10-m wind speed predictions. Results are from 62 forecasts during November 2008–Jan 2009 over the continental United States.

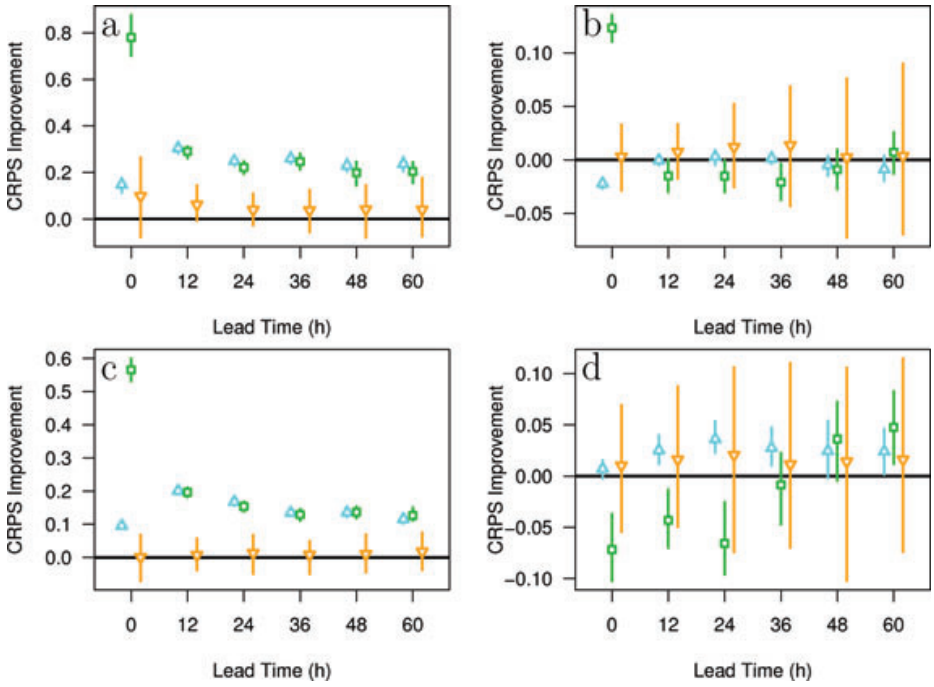

Fig. 6. Continuous rank-probability score difference from *Cntl* for (a) 2-m temperature, (b) 70-kPa temperature, (c) 10-m wind speed and (d) 70-kPa wind speed predictions. Positive indicates an improvement here, and uncertainty in the difference is shown by the vertical lines. Shown are *Phys* (triangle), *PO* (square) and *Param* (inverted triangle). Results are from 62 forecasts during November 2008–January 2009 over the continental United States.

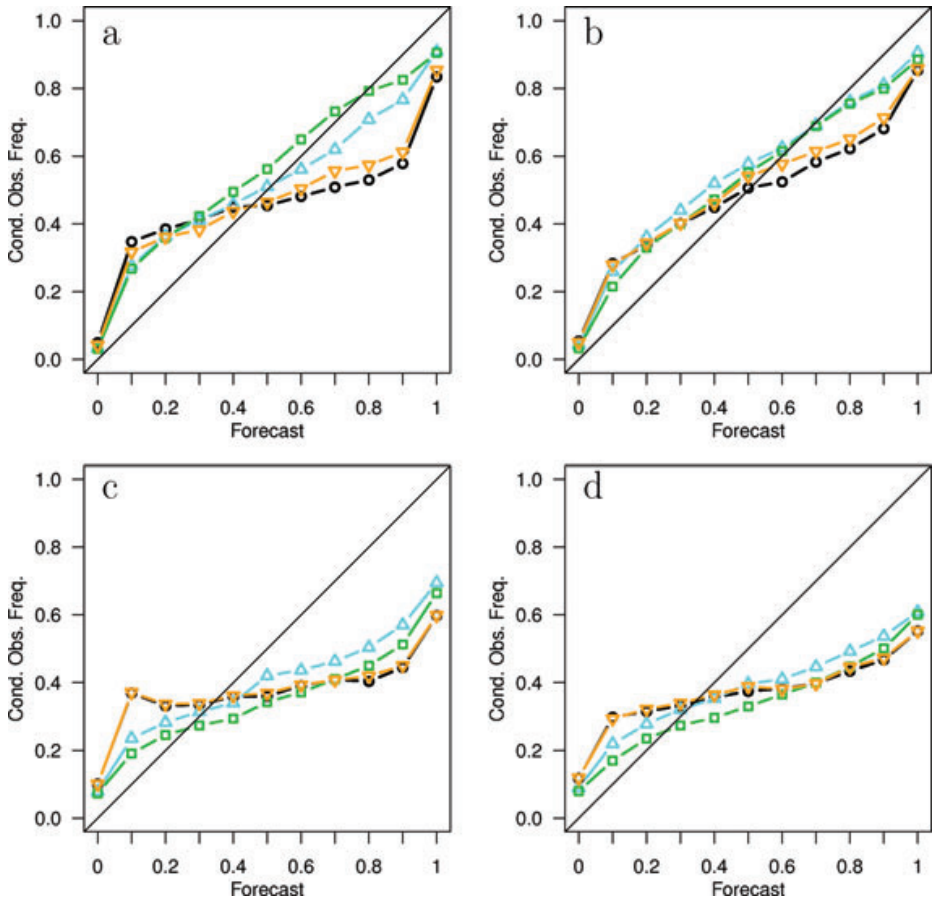

Fig. 7. Reliability diagrams for (a) 12-h 2-m temperature, (b) 48-h 2-m temperature, (c) 12-h 10-m wind speed and (d) 48-h 10-m wind speed predictions. Predictions are for exceeding the 75^{th} percentile of observation distributions. Shown are *Cntl* (circle), *Phys* (triangle), *PO* (square), and *Param* (inverted triangle). Results are from 62 forecasts during Nov 2008 – Jan 2009 over the continental U.S.

probability of exceedance (not shown), thus it is for the less-common events that *Phys* and *PO* stand out most. *Phys* and *PO* also show the greatest reliability for 10-m wind speed (panels c and d), but their distinction is greater for predicting exceedance with low probability. For rarer higher wind events at the surface they perform similarly, and quite poorly. The ensembles are generally more reliable at 70 kPa; there, differences among them are small and we omit those results for brevity. From the lack of differences we can infer that the improved 70-kPa CRPS (Fig. 6) in *Phys* and *Param* results from slightly improved resolution in the ensemble forecasts.

Results presented in this section show that multiple physics suites can improve probabilistic predictions over direct dynamical downscaling of a global ensemble. The use of multiple physics suites appears to improve reliability in particular at the surface, and parameter variations within a single physics suite does not achieve the same result. Given an unbiased forecasts system, which exists here because the biases are explicitly removed, the reliability of an underdispersive ensemble can be immediately improved by increasing the spread. Hacker et al. (2011) shows that the spread of *Phys* is greater than the spread of *Param* at the surface. Although the use of multiple physics schemes introduces logistical complexity to an operational forecast system, we will use *Phys* as the basis for comparison for the remainder of this paper.

4.2. *ETKF and Hybrid*

We turn to the potential for improvement by considering mesoscale IC uncertainty with perturbations introduced on the WRF domain. Ensembles *ETKF* and *Hybrid* are compared to *Phys* for the October 2006 experiment over East Asia (Section 3.2). Both *ETKF* and *Hybrid* use 12-h ensemble forecasts as a prior estimate of analysis perturbations computed via the ETKF algorithm. The primary difference is that *Hybrid* updates the ensemble mean with WRF-3DVar, while *ETKF* perturbations are recentred at each analysis time on the GEFS ensemble-mean analysis. Our goal is to assess the potential for ensemble prediction and we make no attempts to optimally tune the cycling WRF-3DVar (e.g. through improved static background error covariance estimates, choice of weights between static and ensemble-based background error estimates or interval between

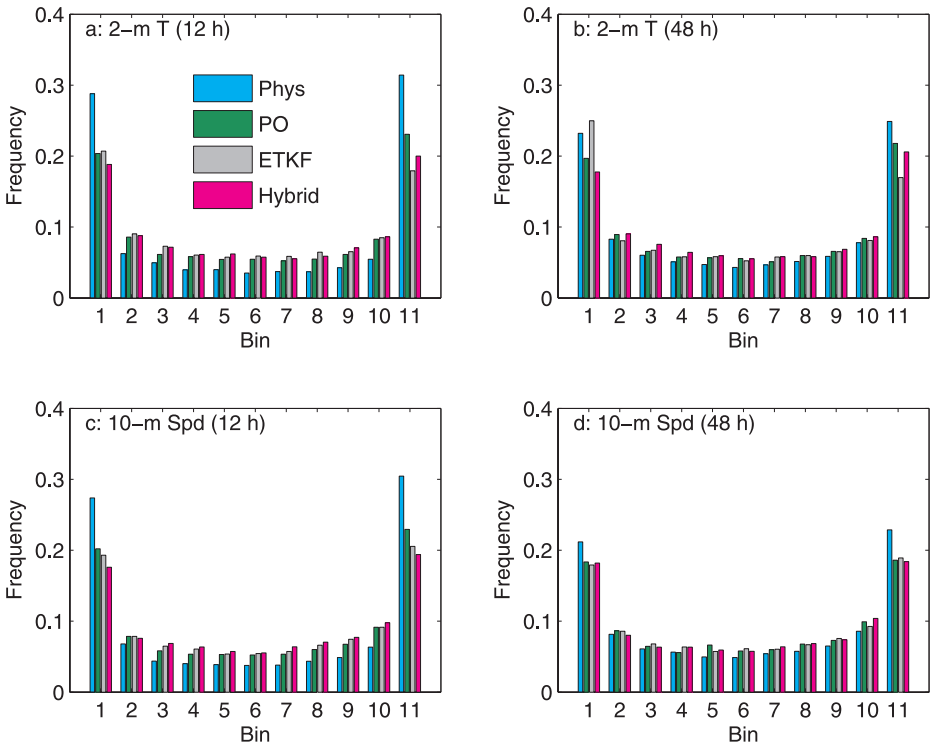

Fig. 8. Rank histograms for (a) 12-h 2-m temperature, (b) 48-h 2-m temperature, (c) 12-h 10-m wind speed and (d) 48-h 10-m wind speed predictions. Results are from 28 forecasts during October 2006 over East Asia.

assimilation updates). If the data assimilation is effective we can expect *Hybrid* to be superior to *ETKF* until LBCs sweep out the influence of the assimilated observations.

Surface rank histograms for 12- and 48-h predictions show a benefit from including mesoscale IC perturbations within the multiple-physics ensemble (Fig. 8). *ETKF* and *Hybrid* are more reliable than *PO*, but the differences diminish by 48 h. At 12-h lead time, *Hybrid* appears slightly more reliable than *ETKF*, but the opposite is arguably true at 48-h lead time. Ensemble dispersion relative to observation variability is also quantitatively similar to the results in Fig. 5, suggesting robustness to season or location.

CRPS differences from *Phys* show that *PO*, *ETKF* and *Hybrid* forecasts are not always an improvement (Fig. 9). Short lead-time improvements at the surface demonstrate the advantages of considering observations, as discussed above with reference to Fig. 6. However those advantages diminish with forecast lead time. Short lead-time CRPS deterioration for *PO* and *Hybrid* 70-kPa wind speed (Fig. 9d) compared to *Phys* occurs over East Asia, as was also observed for *PO* relative to *Cntl* over CONUS (Fig. 6d). Because *Hybrid* and *PO* assimilate data to produce an ensemble mean analysis, while *ETKF* is centred on the GEFS ensemble mean, the results imply that data assimilation is harming skill in the analysis and very short range. It is unclear why degradation appears for 70 kPa winds, but only *PO* shows a (weaker) effect for 70 kPa temperature and those ensembles show improvement at the surface.

5. Stochastic backscatter and limited multiphysics

Results presented above show that multiple physics suites offer an immediate benefit compared to a single-model implementation; this result is consistent with many prior studies. In an effort to reduce development and maintenance complexity and cost, we seek a system that performs as well or better than *Phys* with fewer physics suites. Here we ignore explicit treatment of mesoscale IC uncertainty, and instead rely on large-scale IC uncertainty from the GEFS, as in *Cntl*, *Phys* and *Param* above. Returning to the CONUS winter-time experiment, we find that it is possible to deploy a more skilful ensemble by using stochastic parametrizations, fewer physics suites and within-physics parameter perturbations.

Profiles of total spread and RMSE show that varying physics is beneficial in the lower atmosphere, while stochastic perturbations are most effective above 70 kPa (Fig. 10). Using both approaches, *LMP2_Stoch* shows the greatest statistical consistency. Above 70 kPa, *Stoch* and *LMP2_Stoch* are virtually indistinguishable. Closer to the surface, *Stoch* is slightly less effective at increasing temperature-prediction spread, and wind-prediction *LMP2* spread exceeds *Stoch* spread at 92.5 kPa. The

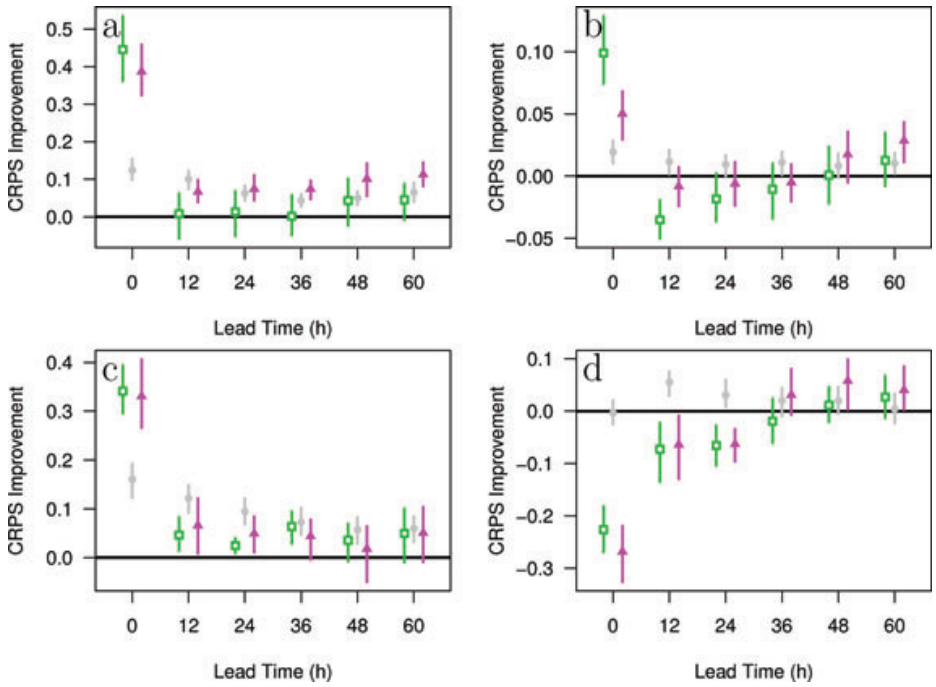

Fig. 9. Continuous rank-probability score difference from *Phys* for (a) 2-m temperature, (b) 70-kPa temperature, (c) 10-m wind speed and (d) 70-kPa wind speed predictions. Positive indicates an improvement here, and uncertainty in the difference is shown by the vertical lines. Shown are *PO* (square), *ETKF* (circle) and *Hybrid* (triangle). Results are from 28 forecasts during October 2006 over East Asia.

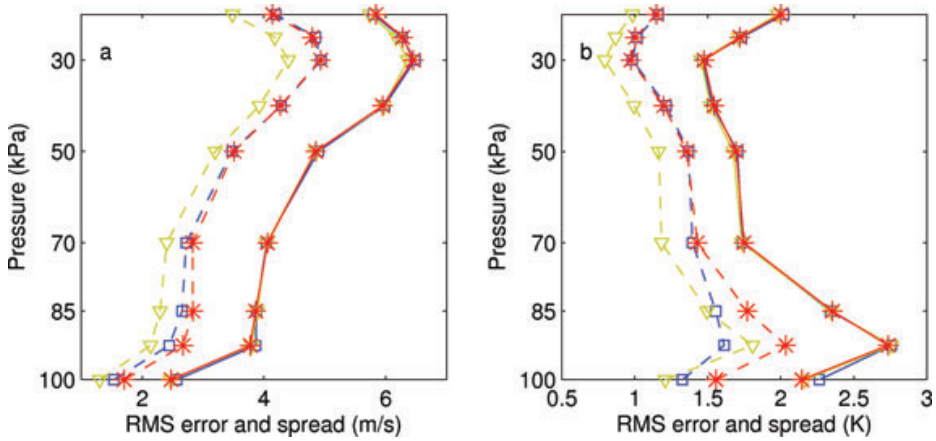

Fig. 10. Root-mean-square ensemble-mean error (RMSE; solid curves) and total spread (dashed curves) of *Stoch* (squares), *LMP2* (inverted triangle) and *LMP2_Stoch* (asterisk). Shown are (a) zonal wind component and (b) temperature for 32 forecasts at 48-h lead time during November–December 2008 over the continental United States.

ensemble-mean RMSE differs only slightly between the ensembles, with a slight advantage to *LMP2_Stoch* for 100 kPa wind predictions.

Near-surface rank histograms give a clearer picture for PBL forecasts (Fig. 11). Temperature and wind predictions from *LMP2_Stoch* at both 12 and 48 h give the flattest rank histograms. The frequency of ranks in extreme bins for 48-h *LMP2_Stoch* predictions is 0.18–0.19, compared to approximately 0.21 for *Phys* and *PO* in Figs. 5(b) and (d). Although the number of forecasts in Figs 5 and 11 is different (62 and 32, respectively), a similar result is found when *Phys* is verified with the same 32 forecasts (not shown).

CRPS differences with *Phys* show that neither *Stoch* nor *LMP2* improve PBL predictions (Figs 12a and c). Because rank histograms indicate comparable reliability, this result suggests slightly deteriorated resolution. Ensemble *LMP2_Stoch* apparently maintains its resolution, or reliability is improved enough to dominate any resolution deterioration. Aloft all of these ensembles show improved CRPS except at analysis time, with *LMP2_Stoch* again demonstrating the greatest skill.

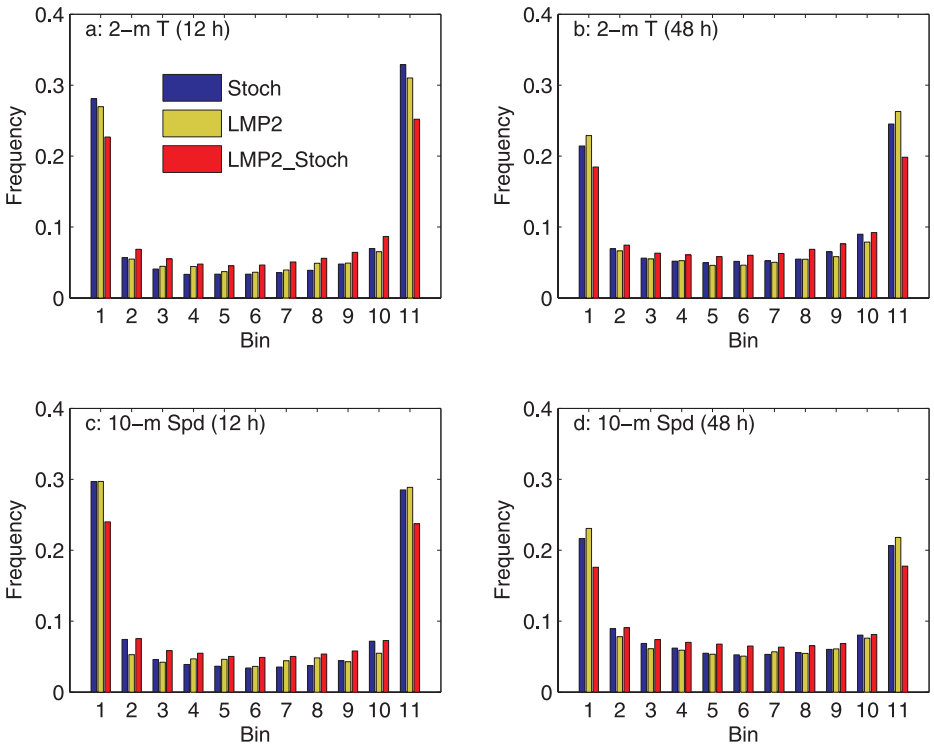

Fig. 11. Rank histograms for (a) 12-h 2-m temperature, (b) 48-h 2-m temperature, (c) 12-h 10-m wind speed and (d) 48-h 10-m wind speed predictions. Results are from 32 forecasts during November–December 2008 over the continental United States.

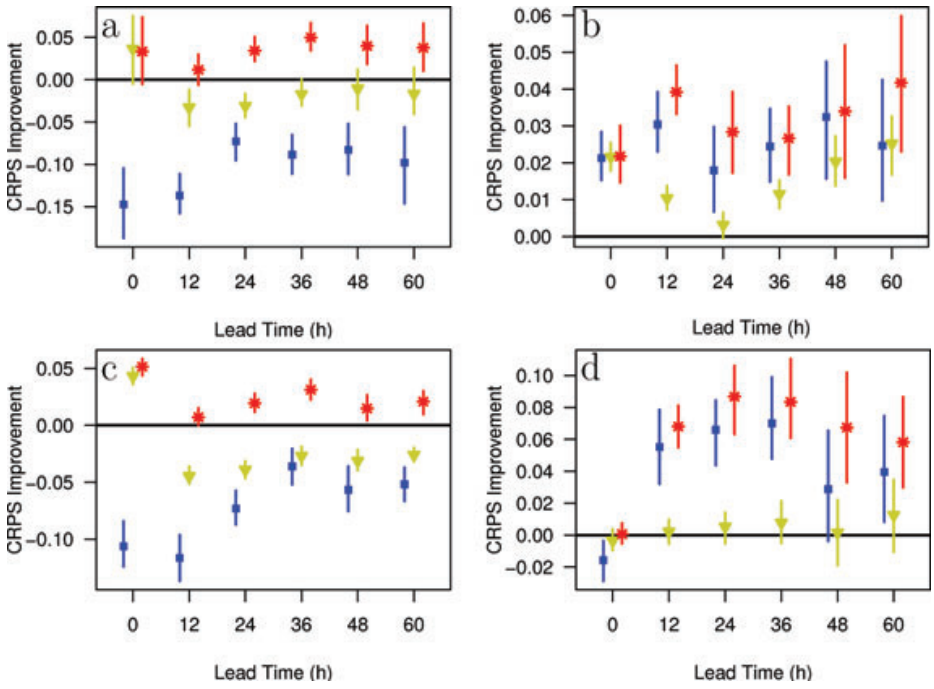

Fig. 12. Continuous rank-probability score difference from *Phys* for (a) 2-m temperature, (b) 70-kPa temperature, (c) 10-m wind speed and (d) 70-kPa wind speed predictions. Positive indicates an improvement here, and uncertainty in the difference is shown by the vertical lines. Shown are *Stoch* (square), *LMP2* (inverted triangle) and *LMP2_Stoch* (asterisk). Results are from 32 forecasts during November–December 2008 over the continental United States.

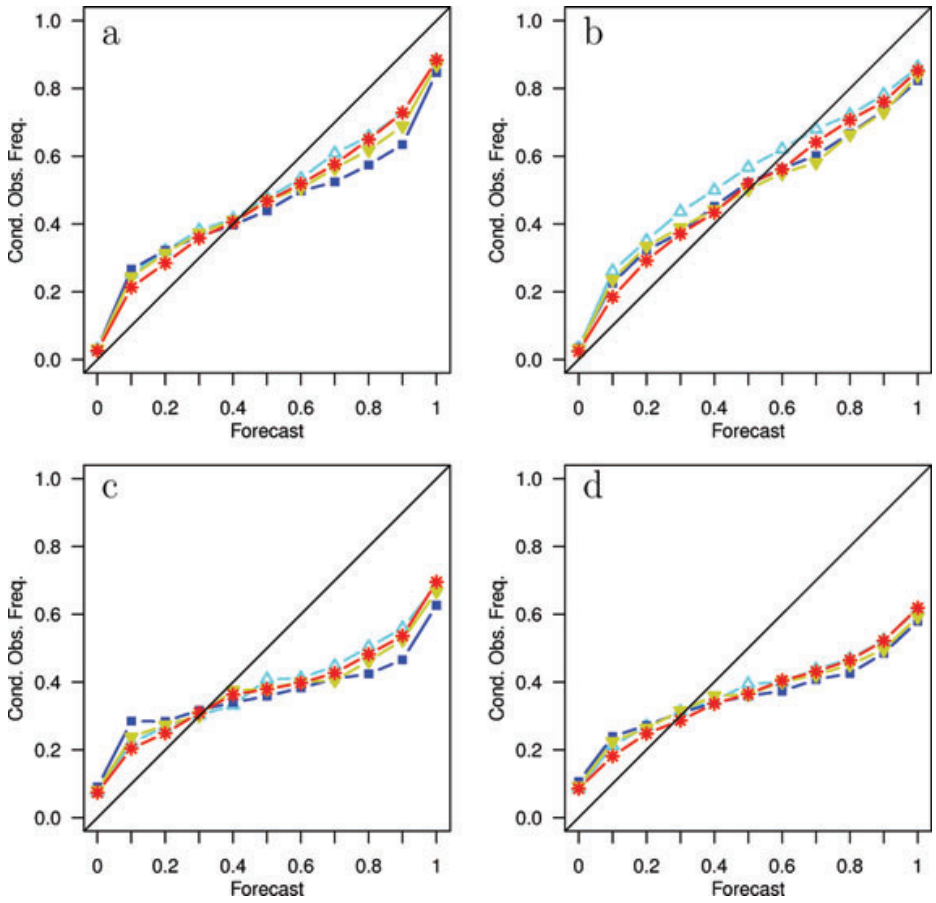

Fig. 13. Reliability diagrams for (a) 12-h 2-m temperature, (b) 48-h 2-m temperature, (c) 12-h 10-m wind speed and (d) 48-h 10-m wind speed predictions. Predictions are for exceeding the 75th percentile of observation distributions. Shown are *Phys* (open triangle), *Stoch* (square), *LMP2* (inverted triangle) and *LMP2_Stoch* (asterisk). Results are from 32 forecasts during November–December 2008 over the continental United States.

Reliability diagrams confirm that *LMP2_Stoch* is the most reliable (Fig. 13). Near-surface wind predictions are quite unreliable for all the ensembles tested in these experiments, but temperature reliability improves during the forecast. Temperature predictions at 70 kPa show high reliability (Fig. 14a), and wind predictions there show greater reliability than at the surface (Fig. 14b).

Relatively poorer skill from *Stoch* in the PBL can be most likely attributed to the fact that the stochastic perturbations were tuned against analyses; near-surface observations were not explicitly considered. Tuning against observations is non-trivial when considering inhomogeneity of the observing network and observation errors. We also should not expect that stochastic perturbations will address much of the model errors in the PBL, which manifest as temporally and spatially varying conditional biases. Additive noise cannot approximate that type of error, but the use of multiple physics suites can apparently capture some of it.

Results in this section reflect the difficult task of producing high-quality ensemble predictions in the PBL. Diversity in physics appears to be the most important ingredient of all those tested here. With some diversity, the skill can be further improved using this stochastic backscatter scheme or introducing parameter variability. In free-tropospheric temperature and wind predictions, stochastic backscatter produces greater skill and statistical consistency than physics diversity; and combining the two leads to still further improvement as discussed in Berner et al. (2011). Details of how these two combine, and in addition combine with parameter perturbations, is left for future study.

6. Summary

This work examines several techniques that attempt to account for mesoscale prediction errors resulting from IC errors and model deficiencies. A practical goal has been to find the most skilful ensemble, with the least degree of complexity, to recommend for implementation in the U.S. Air Force Weather Agency's mesoscale ensemble. Work is ongoing, and this overview also serves as a status report. A summary of findings from these myriad experiments follows:

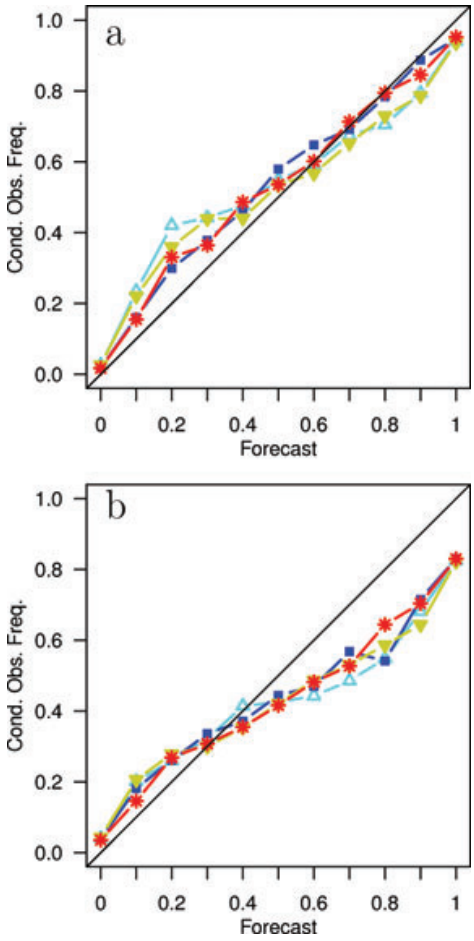

Fig. 14. Reliability of 48-h predictions exceeding the 75th percentile of 70-kPa (a) temperature and (b) wind speed for ensembles *Phys* (open triangle), *Stoch* (square), *LMP2* (inverted triangle) and *LMP2_Stoch* (asterisk). Results are from 32 forecasts during November–December 2008 over the continental United States.

(i) Significant improvement over the direct dynamical downscaling technique of *Cntl* is achievable with all of the methods tested here (Figs 5–7).

(ii) Physics diversity appears critical for probabilistic prediction in the PBL (Figs 10–13).

(iii) All of the ensembles except *Cntl* show improvement over the multiphysics ensemble (*Phys*) for at least one of the metrics examined here. Statistically significant improvement is observed for all ensembles except the multiparameter ensemble *Param* (Figs 5–9, 12). None of the ensembles except *Cntl* are significantly worse than *Phys* under any metric, except for *PO* and *Hybrid* 70-kPa wind CRPS at short forecast ranges.

(iv) Results from *PO*, *ETKF* and *Hybrid*, which take into account IC uncertainty, show IC perturbations are helpful during the first 12 h at the surface in particular. *ETKF* and *Hybrid* are usually more skilful than *PO* (Fig. 9).

(v) When techniques are applied individually, the stochastic backscatter approach shows most skill aloft, and the multiphysics approach shows most skill in the PBL (Figs. 11–13).

(vi) The combination of limited physics variability (three sets), parameter perturbations to those and the stochastic backscatter technique unambiguously results in the best predictions (Figs. 11–13) for these experiments.

(vii) Including observation error estimates in the verification is desirable, but the value of observation errors is unknown and leads to ambiguity in the quantitative skill (Fig. 3). Any calibration approach should consider the best available estimates of observation errors.

Each ensemble technique tested except for *Cntl* represents an explicit attempt to account for mesoscale model or mesoscale IC error. All of the ensembles improve upon *Cntl*, and these methods represent state-of-the-science short-range ensemble prediction. We might conclude then that current systems are somewhat successfully representing mesoscale errors.

Results of these studies also confirm that several outstanding and difficult challenges remain. The extent to which the ensembles are under- or overdispersive will remain unclear unless observation errors are treated properly. Given perfect reliability and statistical consistency, forecast resolution in all quantiles of the observed pdf remains elusive. Calibration can address some of these deficiencies, but calibrating in data-sparse regions is difficult if not impossible. Finally, reducing our dependency on multiple physics schemes (or multiple models) would reduce system complexity and ease interpretation. Because little progress has been made to understand the mechanism by which multiple schemes produces 'good' variability in an ensemble, the community is not yet in a position to learn from multiphysics ensembles and consistently produce comparable or superior ensemble predictions with a single physics suite.

We might speculate why parameter variations appear to be most effective when combined with another model uncertainty approach such as limited multiphysics or stochastic backscatter. Although parameter uncertainty is certainly ubiquitous in NWP models, strongly non-linear and perhaps non-monotonic functions within parametrizations would be needed to effectively change the structure of parametrization output. Intuitively, then, it would be difficult to simulate structural model errors with parameter variations. Although we cannot claim that varying physics schemes correctly simulates model structural error, it at least gives parametrization output with variable structures. Well-constructed stochastic perturbations may also be able to simulate model structural error because it forces the resolved (Reynolds-averaged) model state into phase-space regions it may not otherwise occupy. Parametrization inputs can then be structurally different and result in parametrizations using broader functional ranges. Multiple parameters can account for uncertainty within those otherwise inaccessible neighbourhoods.

Table A1. Physics schemes and parameter perturbations introduced for ensemble *LMP2*.

Scheme	Description	Min.	Default	Max.	Reference
Thompson microphysics	Exponent for raindrop size distribution. 0 recovers exponential.	−0.5	0	0.5	Thompson et al. (2006)
Grell-Devenyi Cu	Number of updraft entrainment and detrainment rates to use in ensemble closure.	3,4	3,3	4,3	Grell and Devenyi (2002)
Mellor-Yamada-Janjić PBL	Background turbulent kinetic energy (squared), used in mixing and determining PBL depth.	0.02	0.2	0.25	Janjić (2001)
Eta microphysics (updated)	Raindrop size distribution intercept.	2e6	8e6	2e9	None found
Betts-Miller-Janjić Cu	Slope of cloud efficiency function. Changes adjustment relaxation time.	0.6	0.7	0.8	Janjić (1994)
Community Atmospheric Model short-wave radiation	Exponent in clear-sky transmittance calculation.	0.95	1.0	1.05	Brieglib (1992)

Note: Key references are provided where possible.

Development of ensemble capability is ongoing at AFWA. While developing operational ensemble capability, AFWA has continued research and development plans to add global ensemble members from other national centres. AFWA also plans to investigate the utility and quality impact of an ensemble Kalman filter, the Bayesian Model Averaging calibration (Raftery et al., 2005) technique, and further refinement of the stochastic kinetic energy backscatter scheme (*Stoch*). With these new methodologies AFWA hopes to improve IC spread and forecast probabilities, leading to a true stochastic characterization of flow-dependent predictability.

7. Acknowledgments

This work was funded by the U.S. Air Force Weather Agency. The authors are grateful for support from D. Gill, J. Dudhia and others in the WRF development group.

Appendix

Here we document the parameters perturbed for ensemble *LMP2*. Hacker et al. (2011) discuss in detail the parameters for the operational physics suite in Member 6 (Table 2). Compared to *Param*, forming *LMP2* introduces six additional physics schemes, and each requires parameter perturbations. Parameters for those schemes were chosen based on literature reviews and sensitivity tests. Working with a mid-latitude winter-storm test case and a single parameter and type of physics scheme (e.g. PBL, Cu, etc.) at a time, the sensitivity testing procedure is summarized as follows: (1) A pair of runs with positive (maximum) and negative (minimum) parameter perturbations to Member 6 were completed, and histograms of diagnostic near-surface differences between the two runs were plotted. (2) A pair of runs with positive and negative perturbations to each new candidate parameter independently were completed, and histograms of near-surface grid-point differences between each perturbed pair were plotted. The first-guess perturbation values were chosen arbitrarily but small. (3) Results from (2) were compared to those from (1). (4) Perturbations and runs were repeated if the difference distributions were qualitatively judged too different from those obtained in (1). This last step is a manual and subjective tuning process. The additional physics schemes and brief descriptions of the associated parameters are given in Table A1.

Once the parameter values were chosen, a coin flip sufficed to assign either a maximum or minimim perturbation value to each member and scheme. The control suite of physics (Member 6) is used four times, and Members 7 and 9 are used three times each, to arrive at 10 ensemble members containing a mix of unperturbed and perturbed parameter values in each member.

References

Berner, J., Shutts, G., Leutbecher, M. and Palmer, T. 2009. A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF Ensemble Prediction System. *J. Atmos. Sci.* **66**, 603–626.

Berner, J., Ha, S.-Y., Hacker, J. P., Fournier, A. and Snyder, C. 2011. Model uncertainty in a mesoscale ensemble prediction system: stochastic versus multi-physics representations. *Mon. Wea. Rev.* In press.

Bishop, C. H. Z. T. 1999. Ensemble transformation and adaptive observations. *J. Atmos. Sci.* **56**, 1748–1765.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B. and Beare, S. E. 2008. The MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.* **134**, 703–722.

Brieglib, B. P. 1992. Delta-Eddington approximation for solar radiation in the NCAR Community Climate Model. *J. Geophys. Res.* **97**, 7603–7612.

Burgers, G., VanLeeuwen, P. and Evensen, G. 1998. Analysis scheme in the ensemble Kalman filter. *Mon. Wea. Rev.* **126**, 1719–1724.

Clark, A., Gallus Jr., W. A. and Chen, T.-C. 2008. Contributions of mixed physics and perturbed lateral boundary conditions to the skill and spread of precipitation forecasts from a WRF ensemble. *Mon. Wea. Rev.***136**, 2140–2156.

Desroziers, G., Berre, L., Chapnik, B. and Poli, P. 2005. Diagnosis of observation, background and analysis-error statistics in observation space. *Quart. J. R. Meteor. Soc.* **131**, 3385–3396.

Eckel, F. A. and Mass, C. F. 2005. Aspects of effective mesoscale, short-range, ensemble forecasting. *Wea. Forecast.* **20**, 328–350.

Grell, G. A. and Devenyi, D. 2002. A generalized approach to parameterizing convection combining ensemble and data assimilation techniques. *Geophys. Res. Lett.* **29**. doi:10.1029/2002GL015311.

Grimit, E. P. and Mass, C. F. 2002. Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Wea. Forecast.* **17**, 192–205.

Hacker, J. P., Snyder, C., Ha, S.-Y. and Pocernich, M. 2011. Linear and nonlinear response to parameter variations in a mesoscale model. *Tellus* **63A**, this issue.

Hamill, T. M., Snyder, C. and Morss, R. E. 2000. A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.* **128**, 1835–1851.

Hamill, T. M., Whitaker, J. and Snyder, C. 2001. Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Wea. Rev.* **129**, 2776–2790.

Hou, D., Kalnay, E. and Drogemeier, K. 2001. Objective verification of the SAMEX '98 ensemble experiments. *Mon. Wea. Rev.* **129**, 73–91.

Houtekamer, P. L. and Derome, J. 1995. Methods for ensemble prediction. *Mon. Wea. Rev.* **123**, 2181–2196.

Houtekamer, P. L. and Mitchell, H. L. 2001. A sequential ensemble Kalman filter for atmospheric data assimilation. *Mon. Wea. Rev.* **129**, 123–137.

Janjić, Z. I. 1994. The step-mountain Eta coordinate model: further developments of the convection, viscous sublayer, and turbulence closure schemes. *Mon. Wea. Rev.* **122**, 927–945.

Janjić, Z. I. 2001. Nonsingular implementation of the Mellor-Yamada level 2.5 scheme in the NCEP meso model, Technical Report 437, National Centers for Environmental Prediction Office Note.

Jolliffe, I. T. and Stephenson, D. B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science Jolliffe, I. T. and Stephenson*, D. B. John Wiley and Sons, Chichester.

Lord, R. J., Menzel, W. P. and Pecht, L. E. 1984. ACARS wind measurements: an intercomparison with radiosonde, cloud motion, and VAS thermally derived winds. *J. Atmos. Ocean. Tech.* **1**, 131–137.

Lorenc, A. C. 2003. The potential of the ensemble Kalman filter for NWP: a comparison with 4D-VAR. *Quart. J. R. Meteor. Soc.* **129**, 3183–3203.

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J. and co-authors. 2004. Quantification of of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772.

Parrish, D. F. and Derber, J. 1992. The National Meteorological Center's spectral-statistical interpolation analysis system. *Mon. Wea. Rev.* **120**, 1747–1763.

Raftery, A. E., Gneiting, T., Blabdaoui, F. and Polakowski, M. 2005. Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.* **133**, 1155–1174.

Santer, T. J. and Williams, B. J. 2003. *Design and Analysis of Computer Experiments Santer, T. J. and Williams, B. J.* Springer, New York.

Shutts, G. J. 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. R. Meteor. Soc.* **612**, 3079–3102.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M. and co-authors. 2008. A description of the advanced research WRF Version 3, Technical Report TN-475, National Center for Atmospheric Research.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N. and co-authors. 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406.

Stensrud, D., Bao, J.-W. and Warner, T. T. 2000. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Wea. Rev.* **128**, 2077–2107.

Stensrud, D. J. and Yussouf, N. 2003. Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.* **131**, 2510–2524.

Thompson, G., Rasmussen, R. and Manning, K. 2006. Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part I: description and sensitivity analysis. *Mon. Wea. Rev.* **132**, 519–542.

Velden, C. S., Hayden, C. M., Nieman, S. J., Menzel, W. P., Wanzong, S. and co-authors. 1997. Upper-tropospheric winds derived from geostationary satellite water vapor imagery. *Bull. Am. Meteor. Soc.* **78**, 173–195.

Wang, X. and Bishop, C. H. 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* **60**, 1140–1158.

Wang, X., Bishop, C. H. and Julier, S. J. 2004. Which is better, and ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Wea. Rev.* **132**, 1590–1605.

Wang, X., Snyder, C. and Hamill, T. M. 2007. On the theoretical equivalence of differently proposed ensemble/3D-Var hybrid analysis schemes. *Mon. Wea. Rev.* **135**, 222–227.

Wang, X., Barker, D. M., Snyder, C. and Hamill, T. M. 2008a. A hybrid WRFVAR-ETKF data assimilation scheme for the WRF model. Part I: observing system simulation experiment. *Mon. Wea. Rev.* **136**, 5116–5131.

Wang, X., Barker, D. M., Snyder, C. and Hamill, T. M. 2008b. A hybrid WRFVAR-ETKF data assimilation scheme for the WRF model. Part II: real observation experiments. *Mon. Wea. Rev.* **136**, 5132–5147.

Wei, M., Toth, Z., Wobus, R. and Zhu, Y. 2008. Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus* **60A**, 62–79.

Wilks, D. S. 2006. *Statistical Methods in the Atmospheric Sciences Wilks, D. S*. second edn Elsevier (London).

Ziehmann, C. 2000. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus* **52A**, 280–299.