Model Uncertainty in a Mesoscale Ensemble Prediction System: Stochastic versus Multiphysics Representations

J. BERNER AND S.-Y. HA

National Center for Atmospheric Research, Boulder, Colorado*

J. P. HACKER

Naval Postgraduate School, Monterey, California

A. FOURNIER AND C. SNYDER

National Center for Atmospheric Research, Boulder, Colorado*

(Manuscript received 17 August 2010, in final form 8 November 2010)

ABSTRACT

A multiphysics and a stochastic kinetic-energy backscatter scheme are employed to represent model uncertainty in a mesoscale ensemble prediction system using the Weather Research and Forecasting model. Both model-error schemes lead to significant improvements over the control ensemble system that is simply a downscaled global ensemble forecast with the same physics for each ensemble member. The improvements are evident in verification against both observations and analyses, but different in some details. Overall the stochastic kinetic-energy backscatter scheme outperforms the multiphysics scheme, except near the surface. Best results are obtained when both schemes are used simultaneously, indicating that the model error can best be captured by a combination of multiple schemes.

1. Introduction

The central concern of numerical weather prediction is to predict meso- and synoptic scales of atmospheric motion as accurately as possible. However, since even small uncertainties in the initial conditions or the prediction model will develop over time to meso- and synoptic-scale errors (Lorenz 1963), the predictability of the detailed weather evolution is limited (Lorenz 1969). An objective way to estimate not only the most likely weather evolution, but also the uncertainty of the forecast, is to run an ensemble prediction system, which provides a probabilistic forecast of the atmospheric evolution.

To account for initial-condition error it is now common practice to start each member of the ensemble system from a slightly different initial condition. Initial conditions are most commonly obtained by attempting to perturb the model in directions that will exhibit maximal error growth, for example, by computing singular vectors (e.g., Molteni and Palmer 1993) or bred vectors (e.g., Toth and Kalnay 1993; Houtekamer et al. 1996). However, even with such initial conditions, ensemble forecasts tend to be underdispersive and underestimate the true uncertainty of the atmospheric evolution (Buizza et al. 2005). This leads to unreliable and overconfident probabilistic forecasts, and in particular to a poor representation of large anomalies such as extreme weather events.

Another major contributor to forecast uncertainty is model error (e.g., from parameter and parameterization uncertainty or altogether unrepresented subgrid-scale processes), and only recently have limited attempts to represent model error in probabilistic forecasts been made (Buizza et al. 1999; Stensrud et al. 2000; Palmer 2001; Eckel and Mass 2005; Shutts 2005; Berner et al. 2008, 2009; Bowler et al. 2008, 2009; Li et al. 2008; Plant and Craig 2008; Teixeira and Reynolds 2008; Palmer et al. 2009; Charron et al. 2010; Tennant et al. 2011). As yet there is no unique method the scientific community has agreed upon, partly due to differing views on the nature of model error.

* The National Center for Atmospheric Research is sponsored by the National Science Foundation.

Corresponding author address: J. Berner, NCAR, P.O. Box 3000, Boulder, C0 80305-3000.
E-mail: berner@ucar.edu

DOI: 10.1175/2010MWR3595.1

Model error might arise from a misrepresentation of unresolved subgrid-scale processes that can affect not only the variability, but also the mean error of a model (e.g., Sardeshmukh et al. 2001; Penland 2003; Palmer et al. 2009). While they are caused by the inability to capture all degrees of freedom of the true atmosphere state, the verdict is still open if the subgrid-scale fluctuations must be included explicitly via a stochastic term, or if it is sufficient to include their mean influence by improved deterministic physics parameterizations.

Palmer (2001) suggests that we use stochastic dynamic models (Epstein 1969) such as cellular automata or Markov models to represent model uncertainty due to improperly represented processes in the deterministic models. Others have promoted physics tendency perturbations (Buizza et al. 1999), multiple physics schemes (Murphy et al. 2004), or parameter variations in the physics packages (Stainforth et al. 2005) to account for parameterization uncertainty. Alternatively, the use of multiple models altogether has been shown to provide reliable probabilistic forecasts (Krishnamurti et al. 2000; Hagedorn et al. 2005; Houtekamer et al. 1996).

One advantage of stochastically perturbed models is that all ensemble members have the same climatology and model bias in contrast to multiparameter, multiparameterization, and multimodel ensembles in which each ensemble member is de facto a different model with its own dynamical attractor. From an operational perspective, multiple models or physics schemes require additional resources, since they all have to be maintained and supported.

A first comparison of the different model-error schemes on seasonal to climatic scales is given in Doblas-Reyes et al. (2009). They found that all model-error schemes—multimodel, multiparameter, and stochastic perturbations—lead to significant improvements over unperturbed single-model systems. They also found that the performance details depend on the forecast lead time, and that for lead times shorter than four months, the multimodel gave best results, followed by the stochastic-physics and perturbed-parameter ensemble. However as Doblas-Reyes et al. (2009) pointed out, the various model-error schemes are implemented into different models, making it impossible to fully disentangle core model performance from model-error scheme performance.

Here we implement a scheme using multiple physics combinations (''multiphysics scheme'') and a stochastic kinetic-energy backscatter scheme into the *same* ensemble system and compare their performance to that of the system without model-error representation. We use the U.S. Air Force Weather Agency (AFWA) Joint Mesoscale Ensemble (JME; Hacker et al. 2011), which is a limited-area ensemble system using the Weather Research and Forecasting model (WRF) with Advanced Research WRF (ARW-WRF) dynamic solver, version 3.1.1. Limited-area ensemble systems focus on subsynoptic time and length scales and are faced with additional challenges such as uncertainties in the lateral boundary conditions, which are also substantial sources of model error. Here, we focus on the ''internal'' model-error component of the forecasting system by using the same initial and boundary conditions for all experiments and evaluate the relative performance of the various schemes.

The idea of stochastic kinetic-energy backscatter of subgrid-scale fluctuations (Mason and Thomson 1992) was originally developed in the context of large-eddy simulation and is based on the notion that the turbulent dissipation rate is the difference between upscale and downscale spectral transfer, with the upscale component being available to the resolved flow as a kinetic-energy source. Shutts (2005) adapted these concepts subsequently to numerical weather prediction.

Berner et al. (2009) report on the performance of a stochastic kinetic-energy backscatter scheme in the ECMWF global ensemble system, and find improved probabilistic skill for medium-range forecasts up to 10 days. The development of this scheme in the ECMWF model is ongoing (for the latest refinements see Palmer et al. 2009). Here, we implement a similar, but simplified scheme into a limited-area model designed for short-range ensemble forecasts of mesoscale events. Hence the focus is on shorter time scales of a few days, and on the performance at the surface and in the boundary layer. Consequently, the different model-error representations are not only verified against analyses, but also observations.

Other variants of Shutts's (2005) stochastic kinetic-energy backscatter scheme have independently been implemented into the Met Office Global and Regional Ensemble Prediction System (MOGREPS) also used for short-range weather forecasting (Bowler et al. 2008, 2009; Tennant et al. 2011) and into the Meteorological Service of Canada (MSC) Ensemble Prediction System for medium-range global forecasts (Charron et al. 2010). These studies also report positive impact on spread, reliability, and probabilistic skill for most of the forecast range. The latest model improvements to the MSC ensemble system include an updated multiparameterization suite as well as two stochastic parameterizations: a stochastic kinetic-energy backscatter scheme and perturbations to the physical tendencies (Charron et al. 2010). They report that the different model-error schemes improve different aspects of probabilistic skill. In particular, the effect of including a stochastic kinetic-energy backscatter scheme is most pronounced in the low-level winds, while using multiple parameterizations for deep convection has a marked positive impact on midtropospheric temperature.

For more theoretical studies of stochastic kinetic-energy backscatter schemes see Frederiksen and Davies (1997, 2004) and Frederiksen and Kepert (2006). This work consists of formulating dynamical subgrid-scale parameterizations based on eddy-damped quasi-normal Markovian, direct-interaction approximation closure models and also a Markov model for the subgrid scale. They found that the kinetic-energy spectra of their large-eddy simulations with subgrid-scale parameterization, including a stochastic backscatter term, agree well with those of the direct numerical simulations.

The goal of the work presented here is threefold: first, we report on the performance of a stochastic kinetic-energy backscatter scheme in a mesoscale ensemble system, with emphasis on both the surface and the boundary layer. Second, the performance of this model-error scheme is compared against a model-error scheme utilizing multiple physics suites within the same ensemble system. Finally, we show how much improvement can be made by combining the two model-error representations. The paper is organized as follows. The experimental setup and data are described in section 2a, section 2b summarizes the multiphysics and stochastic kinetic-energy backscatter schemes, and section 2c defines the verification metrics. The verification of the model-error schemes against observations and analyses is given in section 3, followed by summary and conclusions in sections 4 and 5.

2. Methodology

a. Experiments, data, and verification period

The AFWA JME system is a limited-area ensemble system and uses initial perturbations and lateral boundary conditions from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS; Wei et al. 2008) based on the Global Forecasting System (GFS; Kalnay et al. 1990). Initial conditions in GEFS are generated via an ensemble-transform technique with regional initial-perturbation scaling to account for regional differences in the analysis error variance. Forecasts are made over the conterminous United States (CONUS) domain (see Fig. 1 of Hacker et al. 2011). An overview over the detailed setup of the AFWA mesoscale ensemble and its performance are provided in Hacker et al. (2011). Details on WRF are available in Skamarock et al. (2008).

To represent uncertainties in the land-use characteristics, the AFWA JME system uses perturbed land-use tables, which are generated following a method proposed by Eckel and Mass (2005). Three 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. With those perturbed parameters, different land-use tables are generated for each ensemble member and do not change throughout the experiment [see Hacker et al. (2011), especially their Fig. 2 for a description of the impact].

The ensemble prediction system was run with ten ensemble members every other day for the period between 21 November 2008 and 13 February 2009, producing a total of forty 10-member ensemble forecasts. Forecast are initialized daily at 0000 UTC and integrated for 60 h. All results here use a grid spacing of 45 km in the horizontal and 40 vertical levels. For the subperiod 21 November–21 December 2008, additional forecasts were initialized at 1200 UTC.

It is known that the performance of an ensemble system can depend on the verification reference (e.g., Bowler et al. 2008), that is, if the model is verified against observations or analyses. Since both ways have different advantages and disadvantages as outlined below, here all experiments are verified against both observations and analyses.

The GFS analysis is taken from the NCEP Global Data Assimilation System (GDAS) in which surface observations were assimilated. An advantage is that the analysis is available at each grid point and level; a disadvantage is that it is not generated by the same WRF model as the forecasts and that it is difficult to estimate the analysis error. The evaluation against observations is performed at 106 upper-air sounding stations providing vertical profiles with 11 mandatory pressure levels, and at about 3000 surface stations for the aviation routine weather report (METAR) measurements. A disadvantage of the observation diagnostics is that there are relatively few observation stations, that the data from these stations can have missing values, and that the model output has to be interpolated to the station locations. Observation and analysis errors are a substantial component of forecast error and need to be included in an accurate verification (e.g., to determine if an ensemble system is under- or overdispersive). However, since we have little trust in the estimate of observation and analysis errors (Hacker et al. 2011), we mostly omit them in this study.

Since our interest is on the relative performance of different model-error schemes, rather than on their absolute performance, this omission should not affect our findings.

The verification against analysis is done for the entire period from November 2009 to February 2010, but only for 0000 UTC initializations every other day, resulting in a total of 40 ensemble forecasts. For the verification against observations, initializations for 0000 and 1200 UTC were used, but only for a month from 21 November–21 December at odd days, leading to a total of 30 ensemble forecasts. While the details of the verification period differ

TABLE 1. Experimental setup.

Expt	Physics package/ model-error scheme	Line style
CNTL	Control physics	Black dashed
PHYS	Multiple physics schemes	Gray solid
STOCH	Stochastic kinetic-energy backscatter scheme	Black solid
PHYS_STOCH	Multiple physics schemes and stochastic kinetic-energy backscatter scheme	Gray dashed

somewhat depending on which data are used for the verification, they are similar enough to enable a direct comparison. To confirm this, we repeated all verifications on the intersection of the verification periods. The results were qualitatively the same, but since less data were used, less significant. To maximize the significance of the results presented, we decided to keep slightly different verification periods.

To quantify the impact of different model-error schemes, four ensemble experiments were conducted (Table 1): The first one is the control ensemble (CNTL) where each member uses the same physics packages. A second experiment comprises a multiphysics ensemble (PHYS), where each member uses a distinctly different set of physics suites. The ensemble system STOCH uses the control physics for each ensemble member (the same as in CNTL), but introduces streamfunction perturbations generated by a stochastic kinetic-energy backscatter scheme. Finally, in the experiment PHYS_STOCH the multiphysics scheme is combined together with the perturbations from the stochastic backscatter scheme.

b. Model-error schemes

1) MULTIPLE-PHYSICS SUITES

The multiphysics ensemble tries to account for parameterization uncertainty by using different parameterizations for land surface, microphysics, planetary boundary layer, cumulus convection, and long- and shortwave radiation. Choosing an optimal set of physics suites, and suites with schemes that work well together, is a challenging and time-consuming task. Here, operational and computational considerations partially constrain the selection of physics suites included in PHYS. Aiming to include as much member independence as possible and produce a skillful ensemble, review of the formulation behind the physics schemes in the WRF (Skamarock et al. 2008) led to 20 candidate ensemble members with suites of physics that differ from each other in one or more fundamental ways. Subsequent testing of those candidate ensemble members for a 1-month trial period, to confirm numerical stability and reveal member behavior, reduced the set to 10 members that runs stably and produces reasonable ensemble forecasts for that period (see also the discussion in Hacker et al. 2011). The resulting multiphysics configuration for the ensemble PHYS are summarized in Table 2. Details on the physical parameterization packages can be found in Skamarock et al. (2008).

This approach is not exhaustive, and results in ensemble members that are not necessarily appropriate for all regions or times of year (e.g., using a microphysics scheme lacking ice-phase physics). We adhere to this constraint and accept the inevitable possibility that there might be better multiphysics combinations for that particular period over the CONUS domain, which points to the more general issue of the difficulty to maintain and run multiphysics schemes operationally. To confirm that there are no members that are substantially less skillful than others, we compared the debiased RMS error for each ensemble member (not shown) and could not find any systematic outliers. We point out that the skill in the ensemble PHYS is the result of the different physics packages in combination with the perturbed land-use parameters.

2) THE STOCHASTIC KINETIC-ENERGY BACKSCATTER SCHEME

The stochastic kinetic-energy backscatter scheme aims at representing model uncertainty resulting from interactions with unresolved scales and is based on the notion that the turbulent dissipation rate is the difference between upscale and downscale spectral transfer, with the parameterized upscale component being available to the resolved flow as a kinetic-energy source (Shutts 2005).

The scheme implemented here is a simplification of the stochastic kinetic-energy backscatter scheme of Berner et al. (2009), who demonstrated its ability to improve the performance in the European Centre for Medium-Range Weather Forecasts (ECMWF) medium-range ensemble forecasting system. The simplification consists of assuming a spatially and temporally constant dissipation rate as discussed below. Here, we derive the equations for the full dissipation rate and comment on the simplifying assumptions at the end of this section. The ECMWF ensemble system is a global model and its dynamical core is pseudospectral with streamfunction as one of its prognostic variables. Since the WRF model is a limited-area model and uses finite differences, the basis functions of the stochastic kinetic-energy backscatter scheme were changed from spherical harmonics to 2D-Fourier modes. In this section we briefly give some motivation and background on the adaption of stochastic kinetic-energy backscatter schemes to numerical weather forecasting, and subsequently describe the equations in the 2D-Fourier basis used here. For details of the derivation we refer to Berner et al. (2009) and in particular, their appendix.

TABLE 2. Configuration of the multiphysics ensemble. Abbreviations are Betts-Miller (BM), Community Atmosphere Model (CAM), Kain–Fritsch (KF), Mellor–Yamada–Janjic (MYJ), Rapid Radiative Transfer Model (RRTM), Rapid Update Cycle (RUC), WRF Single-Moment 6-class (WSM6), and Yonsei University (YSU). For details on the physical parameterization packages and references see Skamarock et al. (2008).

Member	Land surface	Microphysics	PBL	Cumulus	Longwave	Shortwave
1	Thermal	Kessler	YSU	KF	RRTM	Dudhia
2	Thermal	WSM6	MYJ	KF	RRTM	CAM
3	Noah	Kessler	MYJ	BM	CAM	Dudhia
4	Noah	Lin	MYJ	Grell	CAM	CAM
5	Noah	WSM6	YSU	KF	RRTM	Dudhia
6	Noah	WSM6	MYJ	Grell	RRTM	Dudhia
7	RUC	Lin	YSU	BM	CAM	Dudhia
8	RUC	Eta	MYJ	KF	RRTM	Dudhia
9	RUC	Eta	YSU	BM	RRTM	CAM
10	RUC	Thompson	MYJ	Grell	CAM	CAM

To calculate a stochastic kinetic-energy source, random streamfunction perturbations $\Psi'(x, y, t)$ and temperature perturbations $T'(x, y, t)$ are introduced, with a prescribed kinetic-energy spectrum. The *effective streamfunction perturbations* $\Psi'(x, y, t)$ are given by

$$\Psi'(x, y, t) = rD(x, y, t)\psi'(x, y, t), \tag{1}$$

where x is the zonal and y the meridional direction in physical space, and t denotes the time. Here we use x and y rather than the global variables λ and ϕ to emphasize that the domain is limited. Here $D(x, y, t)$ is the local, instantaneous dissipation rate, $\psi'(x, y, t)$ is a 2D *streamfunction pattern* with a prescribed kinetic-energy spectrum, and r is the parameter "backscatter ratio." The spatial and temporal characteristics of the perturbation pattern are controlled by expanding the streamfunction pattern $\psi'(x, y, t)$ in spectral space, and evolving each wavenumber as a first-order autoregressive process as described below. If D is constant, then these characteristics will directly transfer to the effective streamfunction perturbations Ψ'. However, if D is a function of space and time, then the spatial and temporal characteristics will be the convolution between $D(x, y, t)$ and $\psi'(x, y, t)$ (see Fig. 2 in Berner et al. 2009). A difference from Berner et al. (2009) is that in the present work we follow the argument of Shutts (2005), who proposes that the energy in the subgrid-scale should be backscattered onto not only u and v, but also T. Hence, we generate temperature perturbations in the same manner as the streamfunction perturbations and add them to the prognostic equation for the temperature. For brevity, we give here only the derivation for $\psi'(x, y, t)$.

Let $\psi'(x, y, t)$ be a 2D streamfunction-forcing pattern and $u'(x, y, t)$ and $v'(x, y, t)$ the corresponding zonal and meridional wind perturbations, respectively, expressed in 2D Fourier space:

$$\psi'(x, y, t) = \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \psi'_{k,l}(t) e^{2\pi i(kx/X + ly/Y)}, \tag{2}$$

$$u'(x, y, t) = -\frac{\partial \psi'(x, y, t)}{\partial y} = -\frac{2\pi i}{Y} \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} l\psi'_{k,l}(t) e^{2\pi i(kx/X + ly/Y)}, \tag{3}$$

$$v'(x, y, t) = \frac{\partial \psi'(x, y, t)}{\partial x} = \frac{2\pi i}{X} \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} k\psi'_{k,l}(t) e^{2\pi i(kx/X + ly/Y)}, \tag{4}$$

where k and l denote the $(K + 1)$ and $(L + 1)$ wavenumber components in the zonal x and meridional y direction in physical space and t denotes time. The Fourier modes $e^{2\pi i(kx/X + ly/Y)}$ form an orthogonal set of basis functions on the rectangular domain $0 < x < X$ and $0 < y < Y$. If the $\psi'_{k,l}$ are nonvanishing for at least one $|k| < K/2$ or $|l| < L/2$ and do not follow a white-noise spectrum, the streamfunction perturbations will be spatially correlated in physical space. The Fourier expansion implies doubly periodic boundaries. This imposes some constraints on the pattern, but since it is only used for perturbations, we do not anticipate any problems.

Since the physical processes mimicked by this streamfunction forcing have finite correlation times, we introduce temporal correlations by evolving each spectral coefficient by a first-order autoregressive (AR1) process:

$$\psi'_{k,l}(t + \Delta t) = (1 - \alpha)\psi'_{k,l}(t) + g_{k,l}\sqrt{\alpha}\varepsilon_{k,l}(t), \tag{5}$$

where $(1-\alpha)$ is the linear autoregressive parameter, $g_{k,l}$ is the wavenumber-dependent noise amplitude, and $\varepsilon_{k,l}$ is a complex-valued Gaussian white-noise process with mean $\langle \varepsilon_{k,l}(t) \rangle = 0$ and covariance $\langle \varepsilon_{k,l}(s)\varepsilon^*_{m,n}(t) \rangle = \sigma^2 \delta_{k,m}\delta_{l,n}\delta_{s,t}$. The * denotes the complex conjugate. In addition, we assume $\alpha \in (0,1]$ (i.e., we exclude the case of a nonfluctuating forcing). The variance and autocorrelation of the AR1 are well-known quantities (e.g., von Storch and Zwiers 1999) and are given for the Markov process in (5) by

$$\langle \psi'_{k,l}(t)\psi'^*_{k,l}(t) \rangle = \frac{g^2_{k,l}\sigma^2}{2-\alpha} \quad \text{and}$$
$$\frac{\langle \psi'_{k,l}(t+\Delta t)\psi'^*_{k,l}(t) \rangle}{\langle \psi'_{k,l}(t)\psi'^*_{k,l}(t) \rangle} = 1-\alpha. \tag{6}$$

Here we interpret (5) as the discrete approximation of a Stratonovitch stochastic differential equation with an exponentially decaying autocorrelation function and a decorrelation time $\tau = \Delta t/\alpha$ (e.g., Berner 2005). The Stratonovitch interpretation is valid for systems where the noise represents continuous processes with decorrelation times smaller than the time increment. For such systems, the noise variance σ^2 and α depend implicitly on the time increment and the $\sqrt{\alpha}$ in front of the noise term guarantees that the noise decorrelates faster than the time step, and fulfills the fluctuation–dissipation relationship. For a detailed discussion see Penland (2003) and references therein.

We furthermore assume that the noise amplitudes follow the power law:

$$g_{k,l} = b\rho^{\beta}_{k,l} \tag{7}$$

with amplitude:

$$b = \left(\frac{\alpha \Delta E'}{4\pi^2\sigma^2\Gamma} \right)^{1/2}, \quad \text{where} \quad \Gamma = \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \rho^{2\beta+2}_{k,l} \tag{8}$$

and $\rho_{k,l} = \sqrt{k^2/X^2 + l^2/Y^2}$ is the effective radial wavenumber. As derived in Berner et al. (2009), this choice of b is such that at each time step Δt a fixed domain-averaged kinetic energy per unit mass:

$$\Delta E' = 2\pi^2 X Y \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \rho^2_{k,l} \langle |\psi_{k,l}(t+\Delta t)|^2 - |\psi_{k,l}(t)|^2 \rangle, \tag{9}$$

$$= 2\pi^2 X Y \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \left(\frac{2}{\alpha} - 1 \right) \rho^2_{k,l} \langle |\psi'_{k,l}(t)|^2 \rangle, \tag{10}$$

$$= \frac{2\pi^2\sigma^2 X Y}{\alpha} \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \rho^2_{k,l} g^2_{k,l}. \tag{11}$$

This is injected into the flow, where the injected energy is given as the difference in the total kinetic energy between time $t + \Delta t$ and time t as expressed by the total streamfunction in Fourier space, $\psi_{k,l}(t)$. In the ensemble-mean sense, these perturbations will inject the domain-averaged kinetic energy $\Delta E'$ given in (11) into the flow.

In summary, a perturbation of the form in (5) with the noise amplitude in (7) will generate streamfunction perturbations with the kinetic-energy spectrum:

$$E_{k,l} = \frac{2\pi^2\sigma^2 X Y}{\alpha} \rho^2_{k,l} g^2_{k,l}. \tag{12}$$

We note that the change of total kinetic energy in (10) does not solely consist of the injected kinetic energy $2\pi^2 X Y \sum_{k=-K/2}^{K/2} \sum_{l=-L/2}^{L/2} \rho^2_{k,l} \langle |\psi'_{k,l}(t)|^2 \rangle$, but is modified by the factor $[(2/\alpha) - 1]$. Berner et al. (2009) show this modification is noise induced and reflects the correlations between the total streamfunction $\psi(x, y, t)$ and the streamfunction forcing $\psi'(x, y, t)$ at time t due to their mutual dependence on the streamfunction forcing at the previous time $t - \Delta t$. If there are no such correlations (i.e., $\alpha = 1$) in the evolution in (5), this factor equals 1 and the change in total kinetic energy equals that of the injected energy assuming that the forcing increments are instantaneously injected at each time step. Second, we remark that if (7) is inserted into the equation for the kinetic-energy spectrum in (12),

$$E_{k,l} = \frac{4\pi^2\sigma^2 b^2}{\alpha} \rho^{2\beta+2}_{k,l}, \tag{13}$$

which states that a streamfunction forcing with power-law $\rho^{\beta}_{k,l}$ will result in a kinetic-energy spectrum of powerlaw $\rho^{2\beta+2}_{k,l}$.

The scheme has a number of parameters that need to be set either based on the best knowledge of the physical processes or by coarse-graining high-resolution model output. The latter was demonstrated by Berner et al. (2009), who used the cloud-permitting model described in Shutts and Palmer (2007) to estimate the power-law exponent $\beta = -1.54$. Following Berner et al. (2009), the other parameters are set to $\sigma^2 = 1/12$ for the noise variance and $1 - \alpha = 0.875$ for the autoregressive parameter, so that each wavenumber has a decorrelation time scale of $\Delta t/\alpha = 0.5$ h, where the model time step is $\Delta t = 240$ s.

The same pattern is used in all vertical levels, leading to a barotropic vertical structure. Experiments with more sophisticated vertical structures (e.g., including vertical correlations from the error covariance matrix of vorticity), did not have a significant impact on the skill of the ECMWF ensemble (unpublished results), which is why we did not improve on this simplification for the present work.

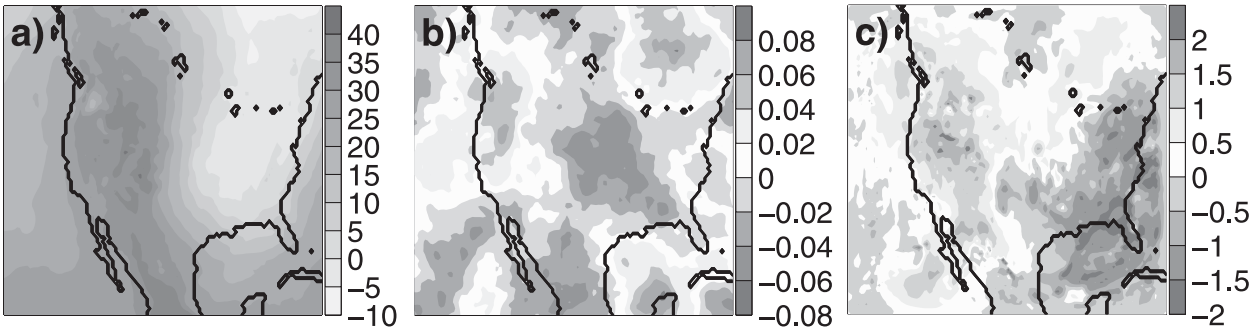

FIG. 1. (a) The 60-h forecast of temperature (°C) at 70 kPa at 1200 UTC 20 Jan 2009 over the conterminous United States for one member of the ensemble system with control physics (CNTL). (b) Snapshot of temperature perturbations from stochastic kinetic-energy backscatter scheme for the same member at a forecast lead time of 59 h. (c) Difference between the original forecast in (a) and the corresponding forecast with the stochastic kinetic-energy backscatter scheme (STOCH) at a lead time of 60 h.

However, future work is planned to develop more favorable vertical structures for the stochastic pattern. Two approaches introducing a vertical structure into the stochastic pattern are described in Palmer et al. (2009) and Tennant et al. (2011).

To obtain a backscatter scheme that is linked to the local instantaneous dissipation rate, Shutts (2005) computes a total dissipation rate $D(x, y, t)$ with contributions from deep convection, numerical dissipation, and gravity/mountain wave drag. Berner et al. (2009) showed that while the flow-dependent formulation of the dissipation rate gave best results, a simplified scheme assuming a spatially and temporally constant dissipation rate D_c led to improvements that were almost as good. Since the correct computation of instantaneous dissipation rates in weather and climate models remains a challenging task, we implement here a simplified scheme assuming a constant backscattered energy rate of $r_\Psi D_c = 2\ \mathrm{m^2\ s^{-3}}$ for streamfunction and $r_T c_p D_c = 2 \times 10^{-6}\ \mathrm{m^2\ s^{-3}}$ for temperature, where r_Ψ and r_T denote the backscatter ratio for streamfunction and temperature, respectively, and c_p is the specific heat. The backscattered energy rates were chosen to yield a reasonable spread and are effectively tuning parameters. A number of short runs with different backscatter rates were conducted to identify values that resulted in the best spread-error consistency without deteriorating the RMS error too much.

The scheme and derivation still use streamfunction as nominal variable, but since WRF uses u and v as prognostic variables, the streamfunction increments $\Delta\Psi'(x, y, t)$ were transformed into u and v increments in (3) and (4) and added to the dynamic equations at end of the dynamical time step, before being passed to the physics routines. The stochastic temperature increments are generated in the same manner as the streamfunction perturbations and added to the prognostic equation for temperature.

An example of the stochastic perturbations in T and their impact on a single forecast valid at 1200 UTC on 20 January 2009 is given in Fig. 1. The 60-h forecast for one member of CNTL is shown in Fig. 1a and its difference from STOCH at the same forecast lead time in Fig. 1c. We see that the largest difference is not necessarily in the regions of the largest perturbation, but in regions with strong gradients where even small perturbations can grow rapidly. This confirms that even by assuming a spatially and temporally constant dissipation rate, the instability of the flow will lead to a flow-dependent error growth.

The impact of the stochastic backscatter scheme for the ensemble forecast valid at 1200 UTC on 20 January 2009, is displayed in Fig. 2 for a lead time of 60 h. Shown are the horizontal pattern of the root-mean-square (RMS) error of the ensemble mean with regard to the analysis at 70 kPa for CNTL and for the spread for three different ensemble systems: CNTL, STOCH, and PHYS. The RMS error of STOCH and PHYS is very similar to that of CNTL and hence omitted here.

As discussed in the next section, the aim is that the spread matches the ensemble mean error in terms of its horizontal pattern and amplitude as closely as possible. We note that the large error in temperature from the south of the Great Lakes to the south-central United States is captured by the spread, but its amplitude is underestimated in CNTL. The spread of PHYS improves on that, but the best representation of the error is given by the spread of STOCH. The patchiness of the RMS error in u is captured well by all ensemble systems; however, they all overestimate the spread in the nonwindy regions. In general, the spread of PHYS is structurally similar but slightly larger than that of CNTL improving the spread-error consistency. The forecast shown in Fig. 2 is a good example of this: the spatial anomaly correlation in the spread between CNTL and PHYS is high, but the maxima of PHYS are slightly larger. STOCH tends to introduce spread in regions not captured by either PHYS

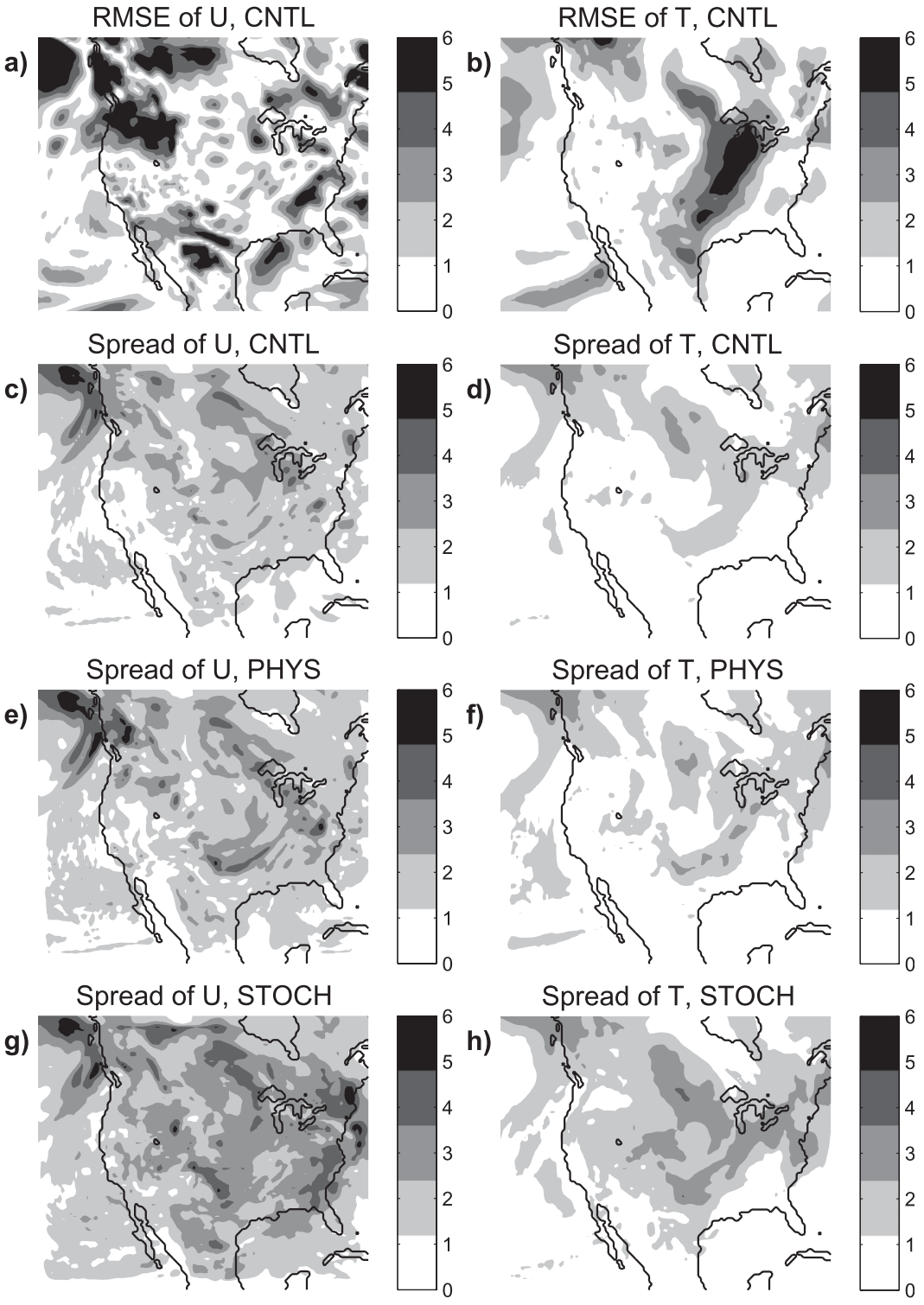

FIG. 2. Horizontal pattern of the RMS error of the ensemble mean with regard to the analysis for the 60-h forecast at 70 kPa at 1200 UTC 20 Jan 2009 in the ensemble system with control physics (CNTL). For (a),(c),(e),(g) variable zonal wind u in m s^{-1} and (b),(d),(f),(h) temperature T in C° are depicted: (a),(b) the RMS error maps; horizontal pattern of the spread in the (c),(d) ensemble systems CNTL; (e),(f) with multiphysics scheme PHYS; and (g),(h) with the stochastic kinetic-energy backscatter scheme (STOCH). For each variable, the gray shadings of error and spread are identical.

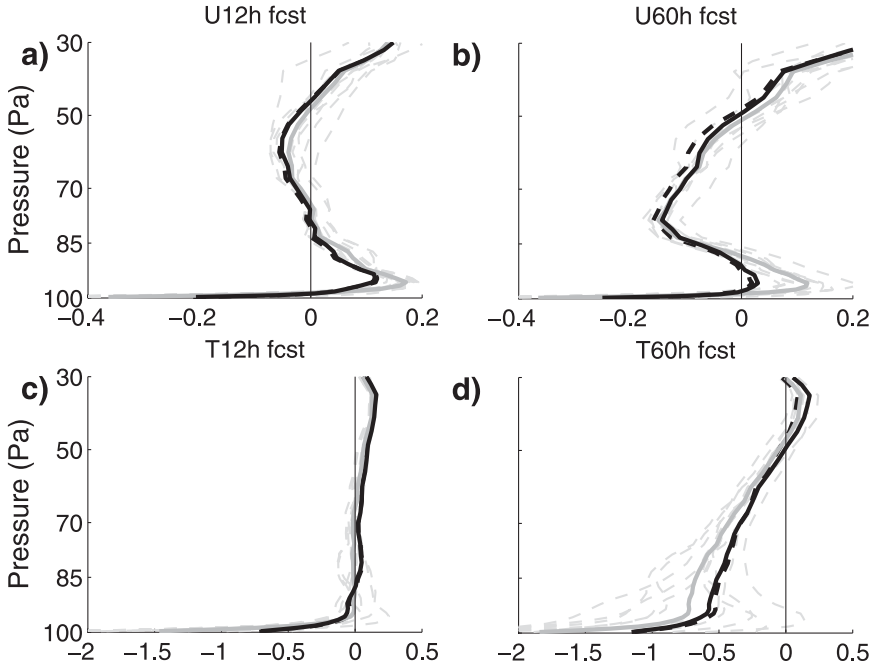

FIG. 3. Mean bias (thick lines) relative to analyses for the winter 2008–09 as function of pressure for (a),(b) u in m s^{-1} and (c),(d) T in K for the ensemble mean of CNTL physics (black dashed), PHYS (gray solid), and STOCH (black solid). The thin gray dashed lines denote the individual member biases of PHYS ensemble. The forecast lead time is (a),(c) 12 or (b),(d) 60 h.

or CNTL. This is sometimes advantageous (e.g., for u along the East Coast) and sometimes not (e.g., over the western United States for T).

c. Metrics for forecast evaluation

The performance of the ensemble systems with and without model-error representation was verified using a number of metrics to assess statistical consistency, reliability, and resolution. The information in different verification metrics is often similar, and so here we present a meaningful subset only.

1) SPREAD-ERROR CONSISTENCY

A measure of reliability is the degree of consistency between ensemble spread and error. A reliable ensemble will exhibit approximate agreement between RMS ensemble-mean error and ''total spread,'' which includes both ensemble spread and observation/analysis error. This approximate agreement expresses the degree to which the ensemble can on average predict the observed or analyzed distribution, and can be expressed as

$$\left[\frac{1}{N-1}\sum_{i=1}^{N}(o_i-\overline{f}_i)^2\right]^{1/2} \approx \left[\frac{1}{N-1}\sum_{i=1}^{N}(\sigma_{f,i}^2+\sigma_{o,i}^2)\right]^{1/2}, \tag{14}$$

where the left-hand side denotes the RMS error of the ensemble mean and the right-hand side the total spread composed of the ensemble variance, $\sigma_{f,i}^2$, and observation (or analysis) error variance $\sigma_{o,i}^2$. The subscript $i = 1, \ldots, N$ indexes the, the total number of verifying observations (or analysis objects) at a particular forecast lead time. Here $\overline{f}_i$ is the ensemble-mean forecast and o_i an observation (or analysis) for this verification time.

2) BRIER SCORE

The performance of the ensemble systems is evaluated using the Brier score as defined in Wilks (1995):

$$\mathrm{BS}(p) = \frac{1}{N}\sum_{i=1}^{N}[g_i(p)-O_i(p)]^2, \quad 0 \le g_i \le 1, O_i \in \{0,1\}, \tag{15}$$

where $g_i(p)$ is the occurrence probability of a dichotomous event E at pressure p for a particular forecast verification date i. If E occurred then let $O_i = 1$; otherwise, $O_i = 0$. From its definition we see, that the better the forecast, the smaller the Brier score and that in the ideal limit of a perfect deterministic forecast BS = 0.

To determine if the performance of the ensemble system depends on the magnitude or sign of the anomalies, the Brier score was categorized in four different verification

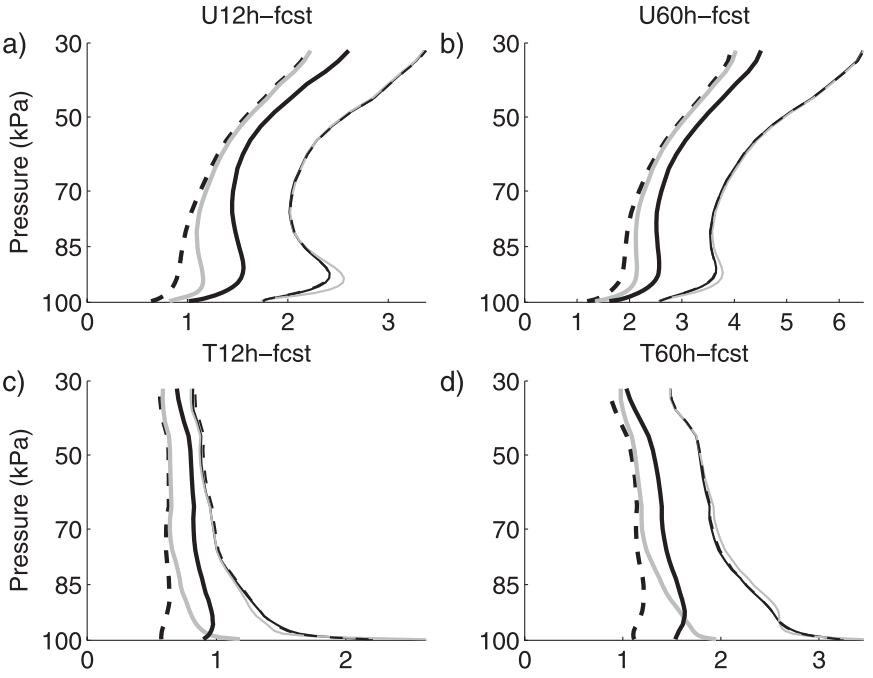

FIG. 4. Spread around ensemble mean (thick curves) and RMS error of ensemble mean relative to analyses (thin curves) for (a),(b) u in m s^{-1} and (c),(d) T in K. Spread and error curves are shown for 3 ensemble systems: control physics CNTL (black dashed), multiphysics PHYS (gray solid), and stochastic backscatter STOCH (black solid). The ensemble systems are debiased with regard to their respective mean monthly bias. Forecast lead time is (a),(c) 12 or (b),(d) 60 h.

events signifying positive or negative "common anomalies" and "extreme events." Here we define common anomalies as an anomaly of less than one standard deviation from the climatological mean, and extreme events as an anomaly of more than one standard deviation. At each isobaric spherical coordinate $\mathbf{r} = (\lambda, \phi, p)$, we compute the climatological standard deviation, $\sigma_x(\mathbf{r})$ of a variable $x \in \{u, T\}$ with regard to their respective monthly mean, and take the weighted average over the number of months in the verification period. Then Brier score profiles as function of pressure level are computed for the four events: $x(\mathbf{r}) < -\sigma_x(\mathbf{r})$, $-\sigma_x(\mathbf{r}) < x(\mathbf{r}) < 0, 0 < x(\mathbf{r}) < \sigma_x(\mathbf{r})$, and $\sigma_x(\mathbf{r}) < x$.

To see if the differences between the schemes are statistically significant, we obtain the empirical distribution of pair-wise Brier score differences by bootstrap sampling with replacement over the N dates. If the difference is positive and statistically significant at the 95% confidence level, we say that model A is significantly better than model B (denoted by a filled diamond marker in the figures). If the difference is negative and statistically significant, this is marked by an open diamond.

3) CONTINUOUS RANK PROBABILITY SCORE

The continuous rank probability score (CRPS) is a generalization of the Brier score to all verification thresholds, and includes contributions from both reliability and resolution. Confidence intervals for the CRPS were obtained in the same manner as for the Brier score differences. For further details on the verification scores, we refer to Jolliffe and Stephenson (2003).

3. Results

The skill of the ensemble systems can be well summarized by the Brier score, the "spread-error consistency," and CRPS, which are the focus of this section. For clarity, the verification against observations and analyses and their respective biases will be discussed separately in sections 3a and 3b. The investigation is centered on the dynamical variables zonal wind u, meridional wind v, and temperature T. However, the results for v are so similar to those of u that all statements and figures made for the zonal wind u are also qualitatively valid for the meridional wind v. Hence, we show no plots for v. The results are discussed for heights below the 30-kPa pressure level.

a. Verification against analyses

For the multiphysics ensemble, each ensemble member has a different physics combination and hence is effectively a different model with its own climatology and bias.

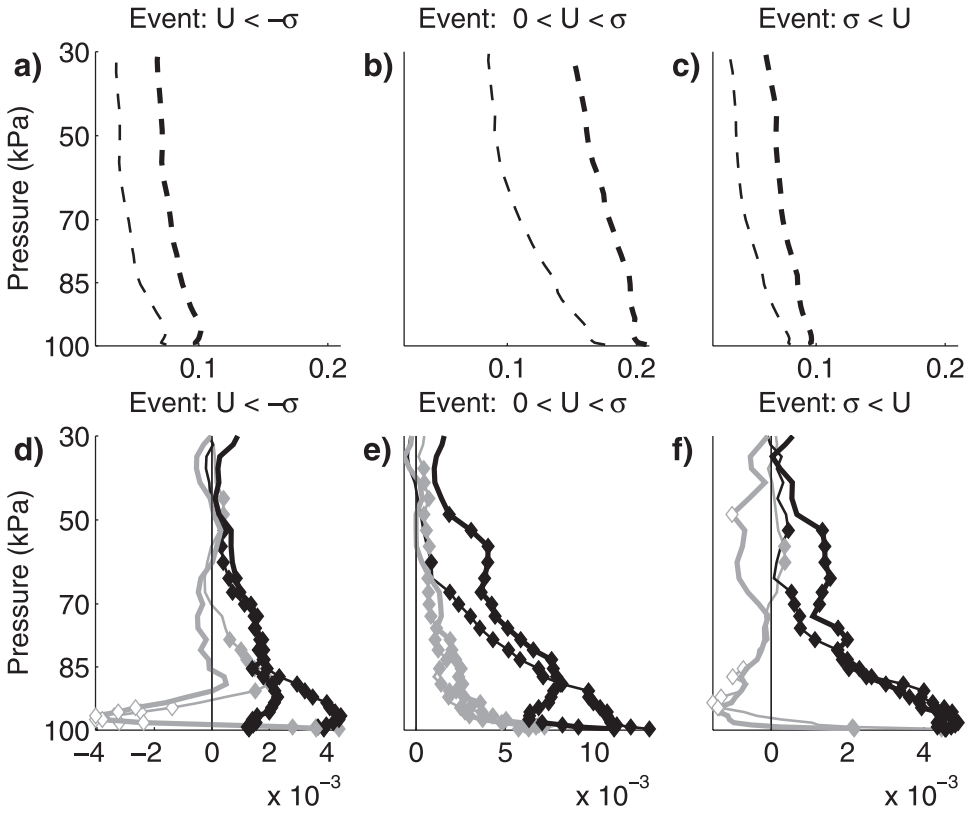

FIG. 5. Brier score of ensemble system with control physics CNTL (black dashed) relative to analyses, for zonal wind u, as function of pressure, for three verification events: (a) $u(\mathbf{r}) < \sigma_u(\mathbf{r})$, (b) $0 < u(\mathbf{r}) < \sigma_u(\mathbf{r})$, and (c) $\sigma_u(\mathbf{r}) < u$; where $\sigma_u(\mathbf{r})$ is the climatological standard deviation of u as function of longitude $\lambda = r_1$, latitude $\phi = r_2$, and pressure $p = r_3$. The score for the event $-\sigma_u(\mathbf{r}) < u(\mathbf{r}) < 0$ is very similar to that in (b) and, hence, not shown. A smaller Brier score denotes better skill. (d)–(f) Brier Score differences of PHYS (gray solid) and STOCH (black solid) from CNTL for the same three events. The sign is defined so that positive differences signify an improvement over CNTL and negative differences a deterioration. Filled (empty) markers denote statistically significant improvement (deterioration) with regard to the skill of CNTL at the 95% confidence level. Forecast lead time is 12 (thin lines) or 60 h (thick lines).

In general, the statistical verification results depend on whether or not the systematic bias is removed prior to the verification. To determine the mean bias for the winter of 2008–09, we compute the bias by subtracting the monthly averaged ensemble mean for each ensemble member from the monthly averaged analysis. Then, we take the mean over the four months of the verification period by weighting each monthly bias with the relative number of dates per month. Subsequently, the horizontal domain average is taken, but the dependence on forecast lead time is kept, since model error evolves with time. For verification purposes the bias was removed a posteriori.

Figure 3 shows the mean bias averaged over all members from each ensemble scheme (thick lines), and the mean biases from individual members in PHYS (gray dashed lines). There are small variances in the member biases even in STOCH and CNTL because of the limited sample size, but they are much smaller than those for PHYS (not shown). At 12-h forecast lead time, the zonal wind is positively biased with the maximum around 90 kPa. At 60 h, the bias at 90 kPa is only evident in PHYS, but all ensemble systems are negatively biased (winds too easterly) farther aloft between 85 and 50 kPa. The largest bias in u is evident at the surface with approximately -0.2 m s^{-1} for STOCH and CNTL and -0.4 m s^{-1} for PHYS showing a large discrepancy between the WRF model and the GFS analysis at the surface.

The temperature bias at 12 h is very small except at the surface where there is a large negative bias of -0.7 K for STOCH and CNTL and -1.5 K for PHYS. At 60 h, the negative bias gets larger, and is now evident in all levels below a height of 50 kPa, which means that the low- to midlevel atmosphere is warmer than the analysis. While the mean biases of STOCH and CNTL are very similar to each other, because of using the same control physics, the mean bias in PHYS is clearly different from the two

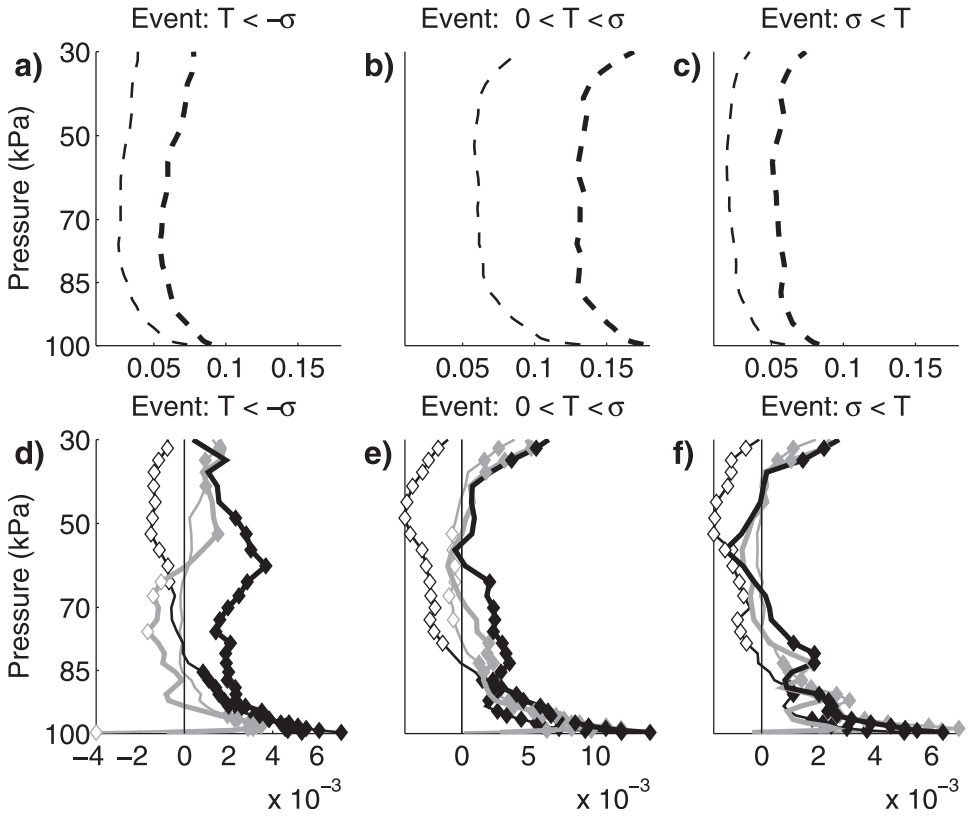

FIG. 6. As in Fig. 5, but for temperature T instead of u.

experiments and gets larger at longer forecast lead times. The mean bias for individual ensemble members in PHYS (gray dashed lines) shows that the systematic bias in the different physics schemes has the largest variability in boundary layer temperature; implying that various combinations of different planetary boundary layer (PBL) schemes and land surface models can produce very different climatologies.

Although a thorough treatment of observed systematic differences is beyond the scope of this work, recent research shows those differences clearly. For example, Santanello et al. (2009) used mixing diagrams to diagnose land–atmosphere coupling for several combinations of PBL schemes and land surface models. They showed that the processes controlling PBL growth and land–atmosphere coupling, including, for example, the Bowen ratio, are integrated in the co-evolution of temperature at 2 m and water vapor, shown by the mixing diagrams. The signature of the mixing diagrams differed according to PBL schemes for dry soils, and according to the land surface model for wet soils.

The spread around the ensemble mean and the RMS error of the ensemble mean are computed for horizontal wind and temperature from all ensemble systems as a function of pressure level at 12- and 60-h forecast lead times, and averaged over the entire domain. Prior to the computation of spread and error, the monthly mean bias (Fig. 3) for each ensemble member was removed.[1] For a perfectly reliable ensemble system the flow-dependent initial uncertainty should be fully represented by the total ensemble spread and thus spread and RMS error should grow at the same rate so that the uncertainty of the forecast is well represented by the ensemble spread. On average, all ensemble systems in this study are underdispersive (i.e., the error exceeds the ensemble spread at all forecast lead times and exceeds it more for longer forecast lead times; Fig. 4).

In the spatial and temporal average of the u-wind component, the error and spread curves have a similar vertical structure with smallest values near the surface, a local maximum near 95 kPa and vertically increasing values above 75 kPa (Figs. 4a,b). For temperature, both the spread and error curves are approximately constant with height for pressures below 80 kPa, and the largest errors near the surface, exceeding 2 K at 60-h forecast lead time (Figs. 4c,d), even after removing the biases. This large discrepancy between model and analyses was also seen in the mean bias (Fig. 3).

[1] To not bias our results in favor of the multiphysics, we removed the bias caused by sampling, in addition to the systematic model bias.

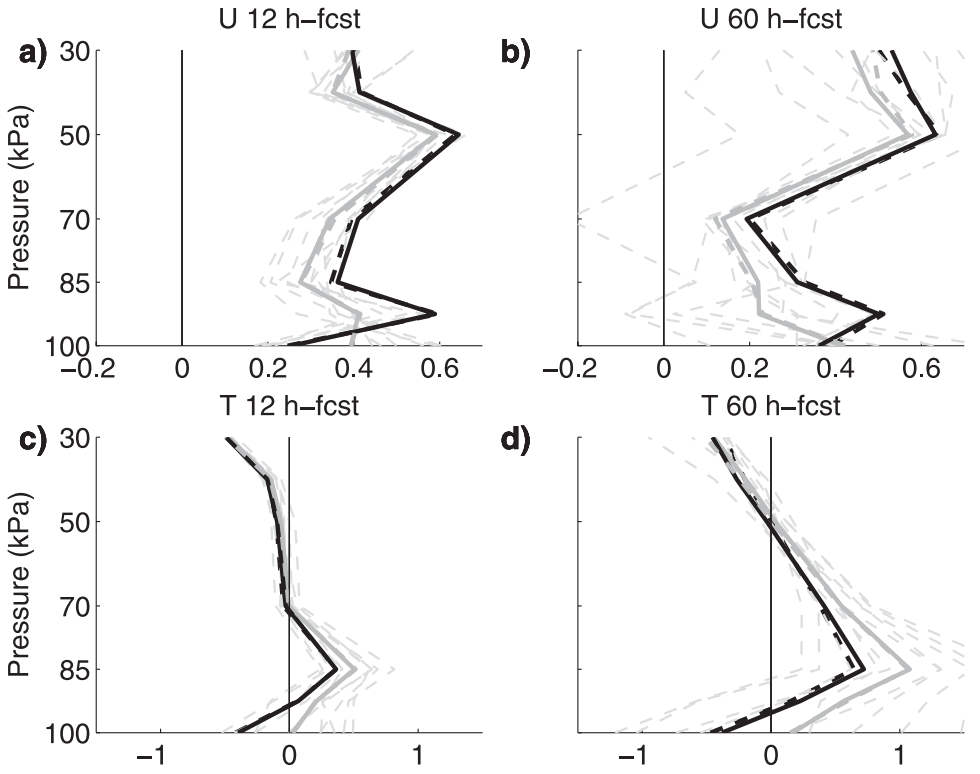

FIG. 7. Mean bias (thick lines) with regard to observations for the winter 2008–09 as function of pressure for (a),(b) zonal wind u in m s^{-1} and (c),(d) temperature T in K for the ensemble mean of CNTL physics (black dashed), PHYS (gray solid), STOCH (black solid), and PHYS_STOCH (gray dashed): (left) 12-h and (right) 60-h forecast lead time. Thin gray dashed lines denote the individual member biases of PHYS ensemble.

We investigated if there are regions that are an exception to the general underdispersiveness. Looking at the temporal averages only, we find that above 60 kPa there are small patches over the ocean and land where the spread is locally larger than the error. However, these overdispersive regions seem random and not linked to any geographical features such as orography. Below 60 kPa the ensemble prediction system is underdispersive everywhere with one exception: there is a small area over the ocean southwest of the Baja peninsula away from the domain boundary, where T is systematically overdispersive.

Comparing the spatially and temporally averaged error and spread curves confirms that the underdispersiveness in the PHYS and STOCH ensembles has greatly improved from CNTL, while the RMS error remains almost the same or is even reduced (Fig. 4). PHYS increases the spread of temperature at 85 kPa by a factor of 1.25 compared to CNTL, and STOCH does so by a factor of 1.5. It is noted that STOCH markedly improves the ensemble spread but hardly degrades the mean errors. Meanwhile, PHYS can increase the mean errors (near 925 hPa in u-wind component) or decrease the mean errors (for the surface temperature) even with smaller increment in the spread. The ensemble mean constitutes the first moment of the ensemble distribution and as such, tends to average over the unpredictable scales of motion. As a second moment, the spread does not have the same filtering qualities. For a linear system we would thus expect that random perturbations would impact the spread, but not the mean. However, in a nonlinear system such as ours, the situation is more complex and it is easy to increase the RMS error by introducing a model error representation. Therefore our findings are nontrivial: while both parameterization diversity and stochastic perturbations lead to an increase in the spread, it is remarkable, that they do so without markedly increasing the RMS error.

In conclusion, the spread-error consistency is improved by both STOCH and PHYS, but the spread generated by the stochastic kinetic-energy backscatter scheme is generally larger than that generated by the multiphysics scheme, without an increase in the mean forecast errors, which results in the most reliable ensemble prediction system.

Verifying the spread-error consistency gives us an estimate of the predictive accuracy of the ensemble-mean forecast. However, the true value of an ensemble system lies in predicting the variability of the potential outcomes. Latter is best assessed by a probabilistic verification. Therefore, we perform such a probabilistic verification

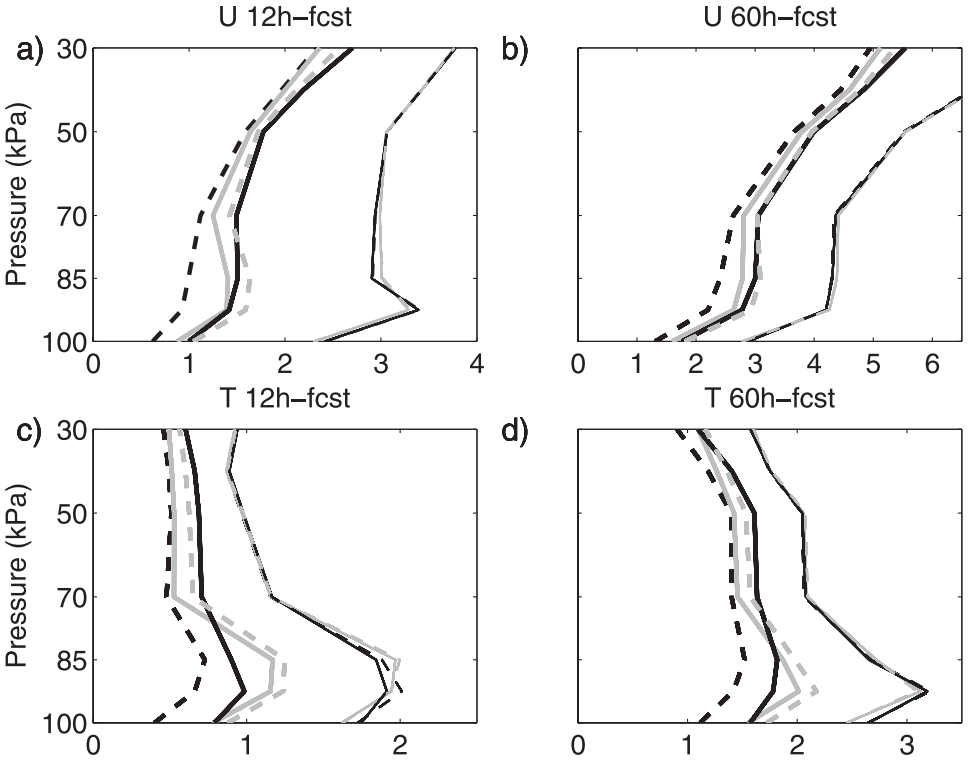

FIG. 8. Spread around ensemble mean (thick curves) and RMS error of ensemble mean (thin curves) relative to observations for (a),(b) u in m s^{-1} and (c),(d) T in K: (left) 12-h and (right) 60-h forecast lead time. Spread and error curves are shown for four ensemble systems: control physics CNTL (black dashed), multiphysics PHYS (gray solid), stochastic backscatter STOCH (black solid), and multiphysics combined with stochastic backscatter PHYS_STOCH (gray dashed). The ensemble systems are debiased with regard to their respective mean monthly bias.

using the most commonly used score, the Brier score. Following common practice, we compute the Brier score after debiasing the monthly mean from each ensemble member. Since the results for the common anomalies, $-\sigma_x(\mathbf{r}) < x(\mathbf{r}) < 0$ and $0 < x(\mathbf{r}) < \sigma_x(\mathbf{r})$, are very similar, we only show plots for the event $0 < x(\mathbf{r}) < \sigma_x(\mathbf{r})$. Both, u wind and temperature forecast time show largest (i.e., worst) Brier scores at the surface, and generally have better probabilistic forecast skills with height [i.e., the scores decrease with height for u (Figs. 5a–c) or remain quasi-constant for heights below the 40-kPa pressure level (Figs. 6a–c)]. The differences between CNTL and the ensembles that vary model-error schemes (Figs. 5 and 6d–f) are generally small compared to the height dependence of the Brier score, but nevertheless tend to be statistically significant at the 95% confidence level (denoted by filled and empty markers) in most levels. The sign is defined so that positive differences signify an improvement over CNTL and negative differences a deterioration. STOCH shows the best skill near the surface and throughout the free atmosphere in the horizontal winds at all lead times (Figs. 5d–f). PHYS and STOCH are both better than CNTL in forecasting the surface temperature, but for short lead times of 12 h, STOCH deteriorates the temperature forecast in heights above 80 kPa (Figs. 6c,d). Interestingly, events more than one climatological standard deviation away from the mean (Figs. 5 and 6a,c) tend to have lower scores and are better captured than are common anomalies, which will be studied further elsewhere. Except that PHYS is statistically significantly worse than CNTL around 925 hPa for strong u-wind forecast (Figs. 5d,f) and STOCH is not as good at predicting T in the free atmosphere at a lead time of 12 h (Figs. 6d–f), both ensembles with model-error representation generally outperform CNTL in most cases at most levels.

b. *Verification against observations*

We have seen that a model-error representation improves the probabilistic forecast when verified against analyses. Next we want to see if this conclusion also holds if we compare the interpolated model output to the observed data. As in the previous section, we look first at the mean bias, followed by a discussion of spread-error

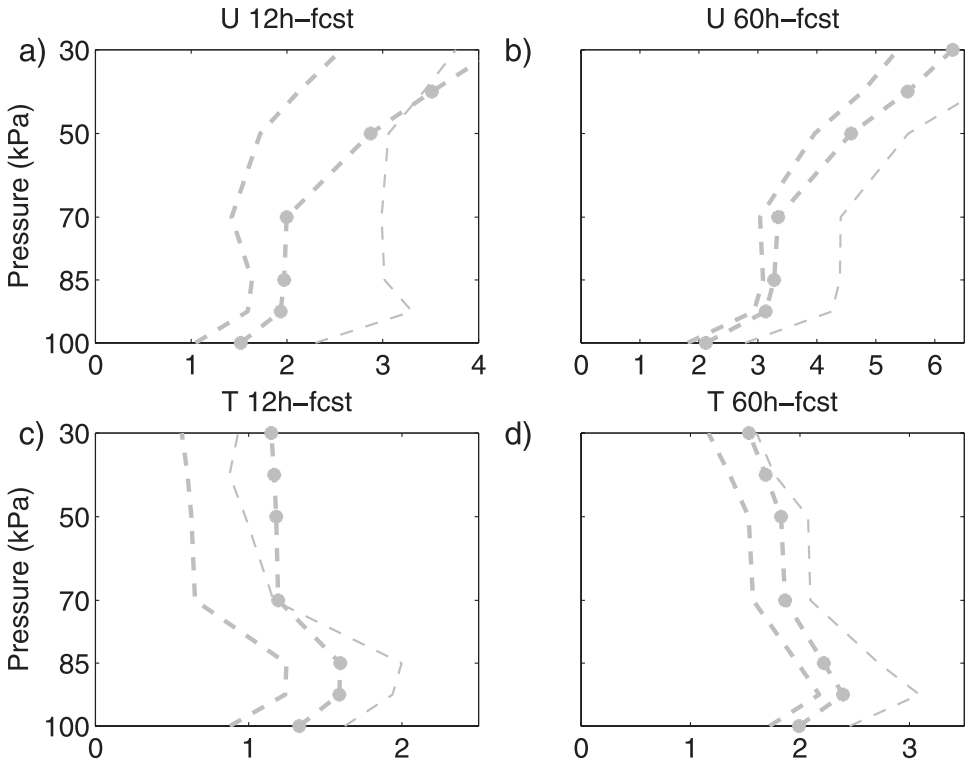

FIG. 9. Spread around ensemble mean (thick curves) and RMS error of ensemble mean (thin curves) for (a),(b) zonal wind u in m s^{-1} and (c),(d) temperature T in K for the multiphysics ensemble PHYS_STOCH. Dashed gray lines without markers denote the debiased ensemble (same curves as in Fig. 8). The spread including our best estimate of observation error is denoted by the thick gray line with markers. Shown are spread and error curves for two different forecast lead times at (a),(c) 12 and (b),(d) 60 h.

consistency and Brier score. For the comparison against observations, results are shown for the additional experiment PHYS_STOCH, which combines the multiphysics and stochastic backscatter schemes (Table 1).

First, we compute the model bias by subtracting the interpolated ensemble forecast from the observations at each station location. Subsequently, the average over the 1-month verification period is taken, for each ensemble member separately. The mean bias relative to the radiosonde observations is mostly positive in the westerly wind at all levels and in temperature in the lower troposphere (Fig. 7). We find that the westerly wind in the forecasts is not as strong as in the observations. The low-level temperature in the ensemble forecast is 0.5 K warmer than the sounding observations, but the analysis shows the boundary layer even colder than the ensemble forecast, especially at 60-h forecast lead time. Considering the common observation errors of 1–1.5 K for surface temperature and of 1.5 m s^{-1} for surface winds, it is found that the ensemble mean forecast is very well matched with the observed climatology within the observational uncertainty. Compared to the observations, temperatures in the ensemble forecasts with control physics (i.e., CNTL and STOCH) are warmer at the surface and colder near the top of the boundary layer. The biases in CNTL and STOCH are quite similar to each other, pointing to the fact that it is the physical parameterizations that are a large contributor to the mean bias. The biases of PHYS and PHYS_STOCH are overall smaller for u except at the surface. For T, the bias of PHYS and PHYS_STOCH is overall larger than for CNTL and STOCH. At the surface, as shown by individual member biases in dashed gray lines, different combinations of surface and PBL schemes produce positive and negative biases resulting in a near-zero mean value.

We note that the profile of the biases verified against observations is quite different from that verified against analyses (Figs. 3 and 7). We have repeated the bias calculation using the intersection of verification periods and found no qualitative change. This confirms that the differences are a reflection of the difference in GFS analysis and observations in combination with interpolation and sampling errors. (There are 106 sounding stations versus $122 \times 98 = 11\,956$ horizontal grid points.) Since the analysis is not produced by WRF, but GFS, we suspect that the bias against the observations is slightly more trustworthy than that against the analysis, especially at

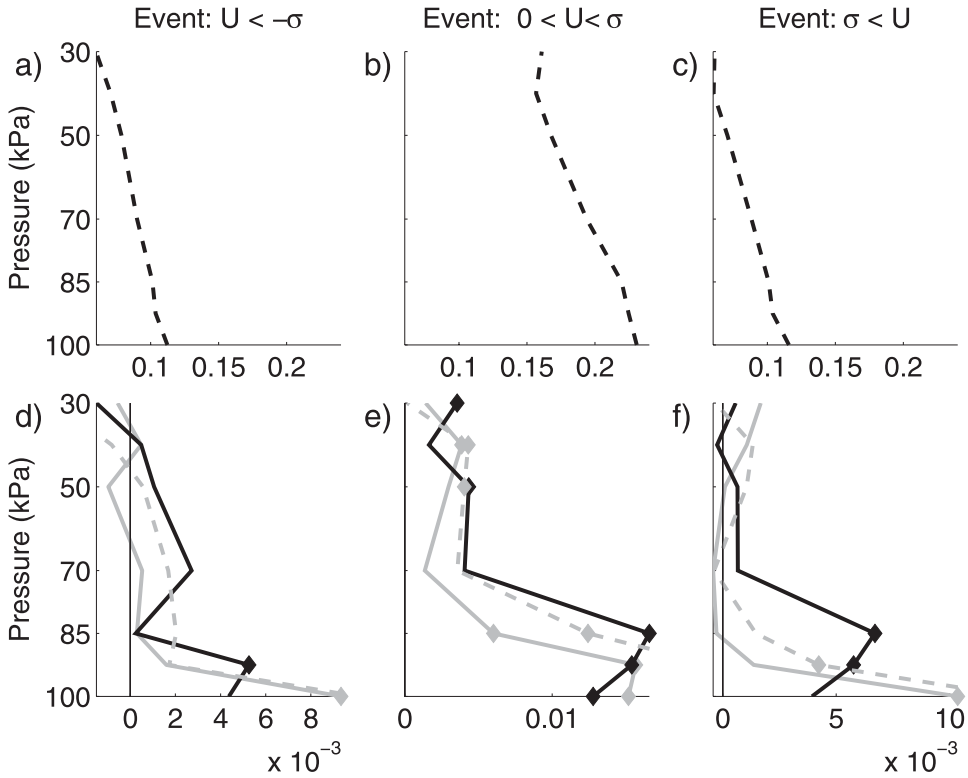

FIG. 10. As in Fig. 5, but for the verification against soundings and for a forecast lead time of 60 h. The Brier score differences of PHYS_STOCH from CNTL are denoted by the gray dashed line.

the surface. On the other hand, by using only 106 sounding stations there is considerable sampling error. Since we debias the data before the verification, our findings will be largely unaffected by the structure of the biases.

The spread and error curves look qualitatively similar to those in section 3a, but there are important differences in the details (Fig. 8). Most markedly, the largest error for T verified against observations occurs now around 925 hPa and not any longer at the surface. Overall, STOCH could best generate the uncertainties in the horizontal winds for the entire atmosphere while PHYS is best at representing the uncertainties in the PBL temperature. The most dispersive ensemble system is PHYS_STOCH, characterized by both perturbations from multiple physics schemes and stochastic perturbations. We note that the combination of both model-error schemes increases the spread in most levels, but not in an additive manner. When comparing PHYS_STOCH to PHYS, we see that the former has considerably more spread, but the RMS error of the ensemble mean is hardly different.

The spread-error consistency indicates again that all ensemble systems considered are underdispersive, and even more so in comparison against analyses. It needs to be stressed that the ensemble appears more underdispersive than it is because we have not accounted for observation error. To get an estimate of the true dispersiveness, Fig. 9 shows spread and error curves for PHYS_STOCH, when the best—albeit unreliable—estimate of observation error has been included. [Note that including the observation error only affects the spread, not the RMS error in (14).] We see that the inclusion of observation error improves the spread-error consistency slightly; however, the ensemble system is still distinctively underdispersive, except maybe for temperature at short lead times and for heights above the 70-kPa pressure level, where the match is quite well.

The Brier score profiles for u and T agree (here shown for at a forecast lead time of 60 h) qualitatively very well with those in section 3a (Figs. 10–11). The verification against observations confirms that both PHYS and STOCH, and their combination perform without exception better than the control ensemble. This is also true for a forecast lead time of 12 h (not shown). Many of these improvements over CNTL are still statistically significant, but because of the smaller sample size, the results are less significant than those presented in Figs. 5 and 6. For u, STOCH performs better than PHYS, especially aloft. For T, the superiority of the stochastic physics ensemble is not as evident as for u especially in the boundary where—in contrast to the previous results—PHYS outperforms STOCH for all events. In the boundary layer and at the surface, the best

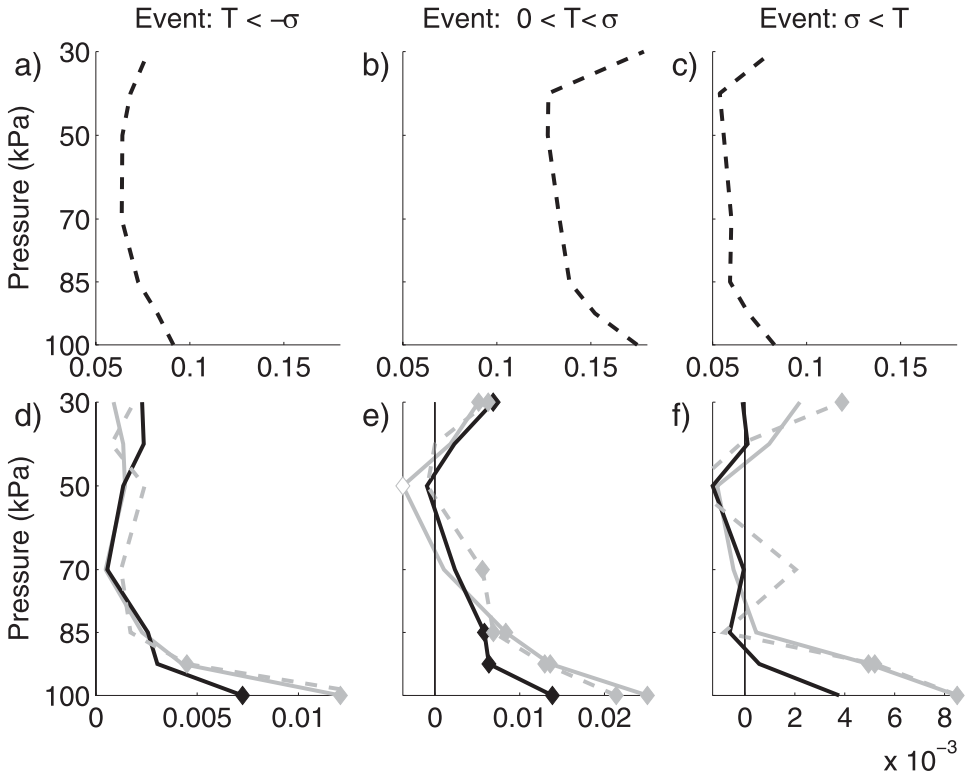

FIG. 11. As in Fig. 10, but for temperature T instead of u.

ensemble system is clearly PHYS_STOCH. It significantly outperforms both PHYS and STOCH for both u and T.

Short-range ensemble prediction systems focus on the mesoscale and are intended to provide accurate near-surface predictions. Since both soundings and analyses have their largest errors in the lowest levels it makes it difficult to determine which results to trust more. Hence, we investigate the performance of the model-error schemes further by verifying against the dense METAR observation network with approximately 3000 stations over the conterminous United States. The focus will be in the temperature at 2 m (T2m) and wind speed at 10 m. As skill score we choose the CRPS, which is a generalization of the Brier score (see section 2c). A comparison of pair-wise CRPS difference at different forecast lead times ranging from 12 to 60 h allows us to examine the relative performance of the ensemble systems (Figs. 12–15). Since CRPS is negatively oriented, we reverse the difference so that improvements over CNTL (top row), PHYS (middle row), or STOCH (bottom row) are shown as positive values. Confidence intervals for the score differences are obtained in the same way as for the Brier score. The intervals depicted as bars denote the 5th and 95th percentiles of the scores differences (see section 2c).

The CRPS results show that at the surface, both model-error schemes outperform CNTL for all forecast lead times (Figs. 12–13, top row). The scores of PHYS and PHYS_STOCH are very similar, but a close investigation shows that PHYS_STOCH is slightly better than PHYS, which is consistent with the Brier score results (Figs. 10 and 11). For the surface variables T2m and wind speed at 10 m, the combination of both model-error schemes, PHYS_STOCH, yields the best performing ensemble system, closely followed by PHYS and then STOCH (Figs. 12 and 13, right column). All score differences are highly significant.

When we investigate continuous rank probability score differences at 70 kPa (i.e., in the free atmosphere), the qualitative performance of different ensemble systems changes somewhat. Note, that in this level the observations consist of upper-air sounding stations. Again all model-error schemes clearly outperform CNTL except at initial time. However, now STOCH is the most skillful ensemble, followed by PHYS_STOCH, and then PHYS (Figs. 14 and 15). The results are again statistically significant for most lead times. An exception is the wind speed at 70 kPa, where the difference in CRPS for PHYS_STOCH and STOCH is not significant.

4. Summary

To account for model uncertainty we tested the performance of three model perturbation schemes: a scheme

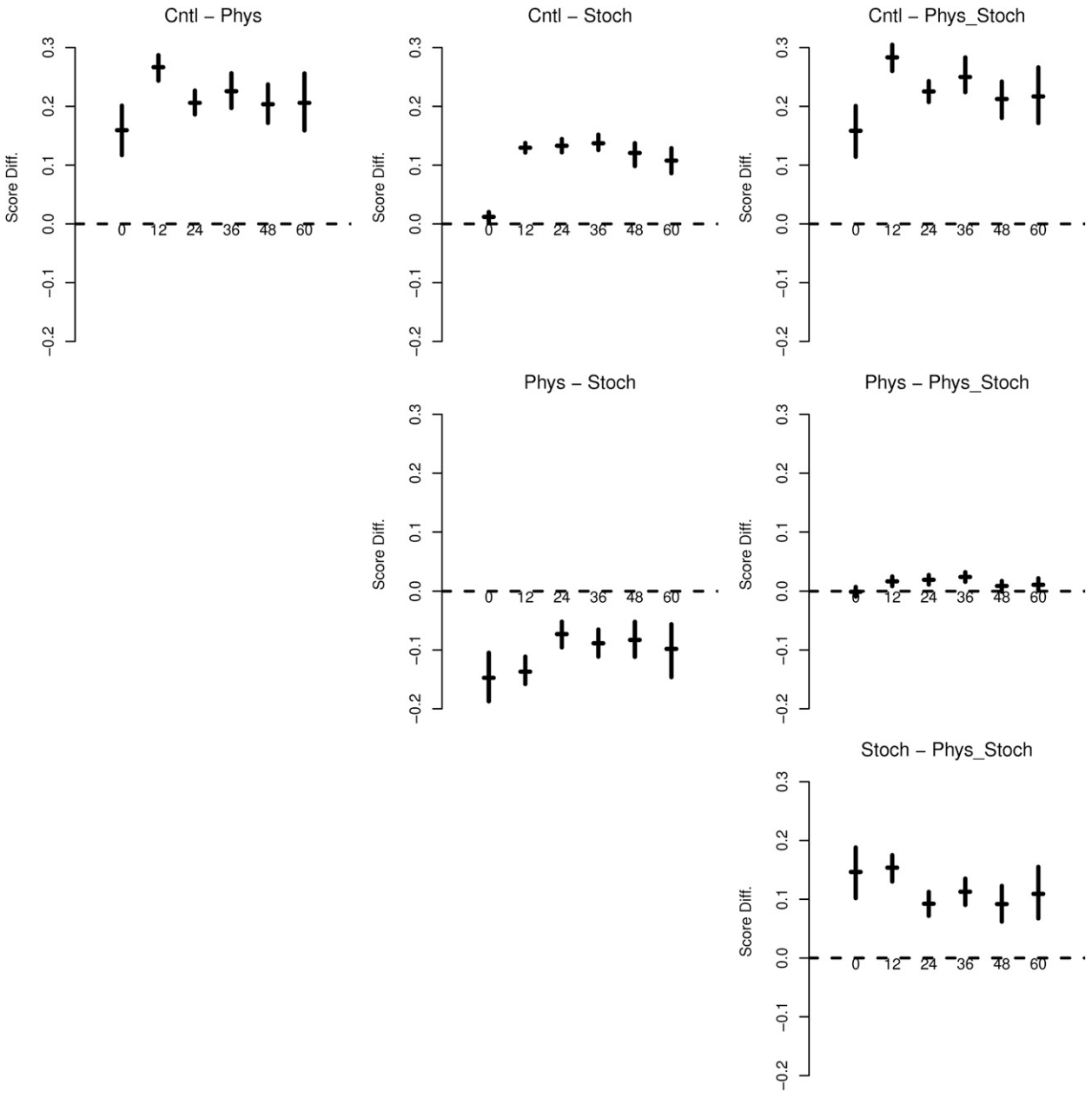

FIG. 12. Pair-wise CRPS difference for temperature at 2m (T2m) for CNTL, PHYS, STOCH, and PHYS_STOCH verified against surface observations. The sign of the differences is defined in such a way that improvements of model A over model B are shown as positive values, where model A is (left) PHYS, (middle) STOCH, or (right) PHYS_STOCH and model B is (top) CNTL, (middle) PHYS, or (bottom) STOCH. The forecast lead times are 0, 12, 24, 36, 48, and 60 h.

using multiple physics suites (PHYS), a stochastic kinetic-energy backscatter scheme (STOCH), and their combination (PHYS_STOCH). The model-error schemes were implemented into the same mesoscale ensemble prediction system and constrained by the same initial and boundary conditions, which allowed for a relatively clean testing of the different model-error schemes.

We found that the qualitative performance of the model is not sensitive to the skill measure used, and hence we focused on the Brier score. For completeness, the model-error schemes were evaluated against both observations and analyses. Overall, the results agree very well, although they are verified over slightly different verification periods—3 months from 21 November 2008 to 13 February 2009 for the comparison with analyses and 1 month from 21 November to 21 December 2008 for the comparison with observations. One exception is the surface, where both analysis error and observation error from soundings are

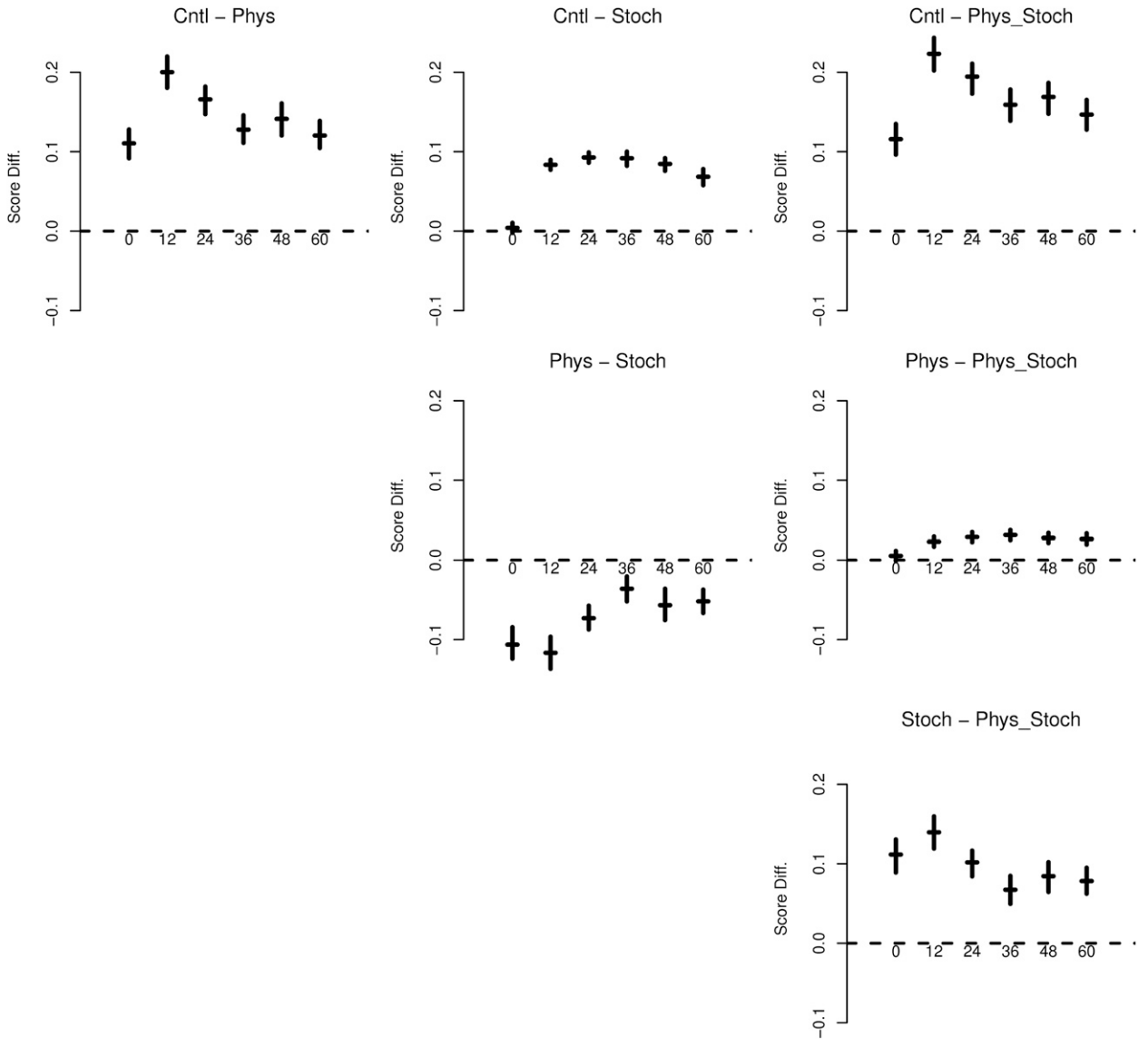

FIG. 13. As in Fig. 12, but for wind speed at 10 m instead of T2m.

large, and the relative merit of the model-error schemes differs somewhat depending on the reference verification.

To summarize the performance, Fig. 16 consists of a pair-wise comparison of Brier scores for a verification against observations. Shown are the variables u, v, and T for forecast lead times of 12 and 60 h, 7 vertical levels and for 4 thresholds for positive and negative small and large anomaly events with regard to the respective climatologies. This amounts to a total of 3 variables $\times$ 2 forecast lead times $\times$ 7 levels $\times$ 4 thresholds = 168 outcomes. Since small Brier scores signify a better forecast, values below the diagonal signify that the model on the ordinate performs better than that on the abscissa. The pair-wise comparison of the number of outcomes (and percentages) of how often model A performs better than model B is given in Table 3, together with the number of outcomes where the score differences are statistically significant. All model-error schemes clearly outperform the control ensemble system with no model-error representation: they have a Brier score better than that of CNTL at least 82% of the time. STOCH performs slightly better than PHYS: 63% of the Brier scores are lower (better) in STOCH than in PHYS. The best-performing ensemble system, PHYS_STOCH, is obtained by combining both model-error schemes. It beats PHYS 79% of the time and STOCH 58% of the time. The same conclusion is drawn when the forecasts are verified against analyses, where STOCH outperforms PHYS in 76% of the cases (Table 4).

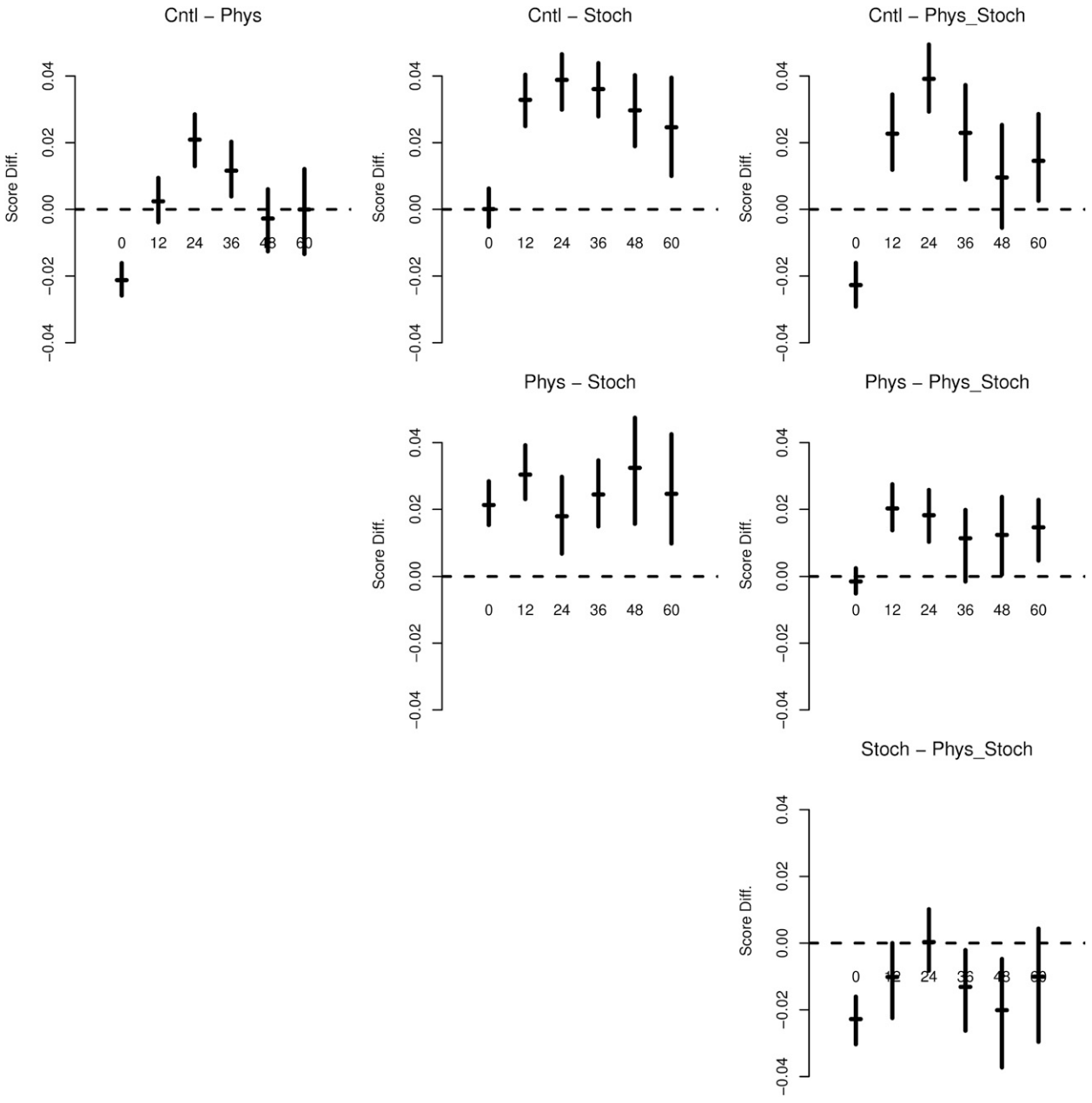

FIG. 14. As in Fig. 12, but for temperature T at 70 kPa instead of T2m. Verification is against soundings.

5. Conclusions

The major findings of this study are the following:

- Including a model-error representation leads to ensemble systems that produce significantly better probabilistic forecasts than a control physics ensemble that uses the same physics schemes for all ensemble members.
- Overall, the stochastic kinetic-energy backscatter scheme outperforms the ensemble system utilizing multiple combinations of different physics-schemes. This is especially the case for u and v in the free atmosphere. However, for T at the surface the multiphysics ensemble produces better probabilistic forecasts, especially when verified against observations.
- The best-performing ensemble system is obtained by combining the multiphysics scheme with the stochastic kinetic-energy backscatter scheme. The superiority of the combined scheme is most evident at the surface and in the boundary layer.
- There is no obvious superiority of one model-error scheme with regard to ''extreme events'' in the sense of anomalies that are greater or smaller than one

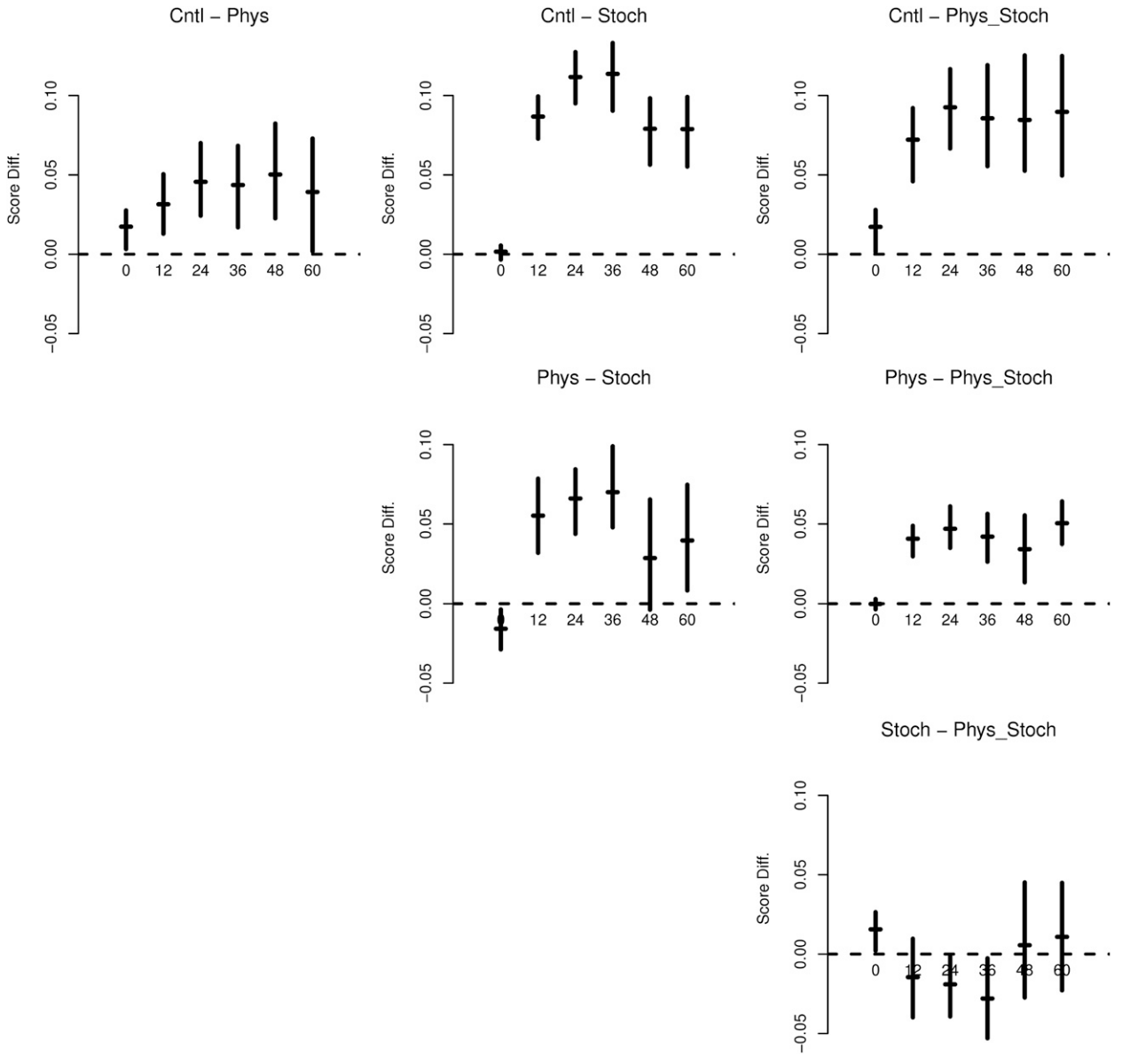

FIG. 15. As in Fig. 12, but for wind speed at 70 kPa instead of T2m. Verification is against soundings.

climatological standard deviation. Rather it seems that an increase in spread helps the score for all thresholds.

- Even with model-error schemes and accounting for observation error, the ensemble spread is still underdispersive. This is consistent with our finding that in general the most dispersive ensemble system is the most skillful.

We conjecture that the superiority of stochastic kinetic-energy backscatter scheme in the free atmosphere occurs because it perturbs the dynamic state directly. The dynamical variables are then fed into the physical parameterizations, which respond to this slightly perturbed dynamical state. This is very different from perturbing the physical tendencies directly, as described in Buizza et al. (1999), which can introduce inconsistencies between the physics and dynamics. The tendency of the model might be to readjust any such inconsistencies at the next time step, possibly leading to erroneous phenomena (e.g., spurious gravity waves). Near the surface, however, the ensemble systems are even more underdispersive than aloft, especially for temperature. Currently, the multiphysics ensemble is more efficient at introducing boundary layer temperature dispersion resulting in better probabilistic skill. Work is planned to improve the performance of the backscatter scheme in the boundary layer.

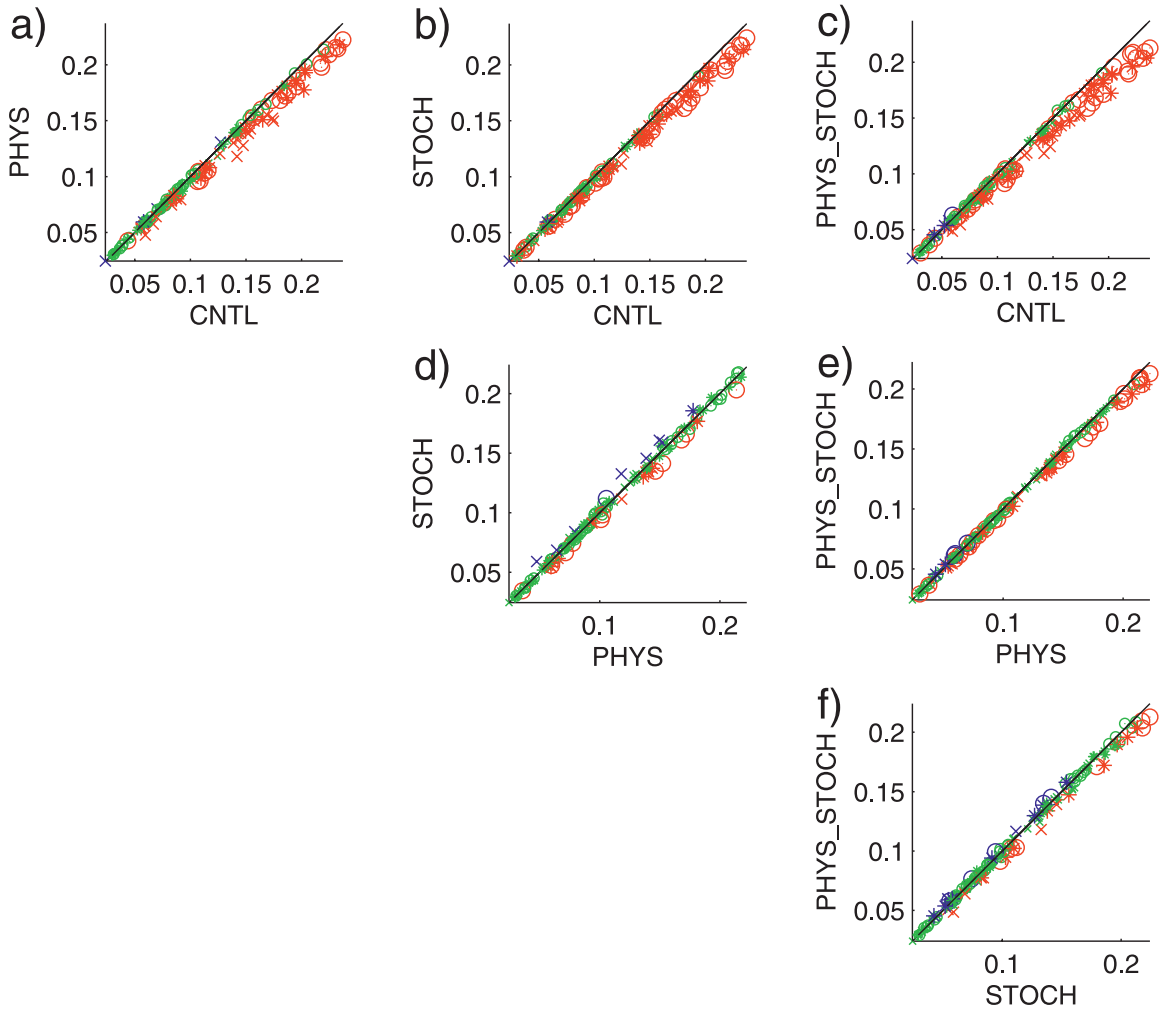

FIG. 16. Pair-wise comparison of the outcomes where model A (ordinates) performs better or worse than model B (abscissas) as measured by the Brier score when verified against observations. Model A is (a) PHYS; (b),(d) STOCH; or (c),(e),(f) PHYS_STOCH and model B is (a)–(c) CNTL; (d),(e) PHYS; or (f) STOCH. Values below the line $BS_A = BS_B$ denote an improvement of model A over B. The outcomes comprise the forecast lead times 12 and 60 h, 4 verification events (see text) and 7 vertical levels for the variables zonal wind u (crosses), meridional v (stars), and temperature T (circles) totaling 168 outcomes. Statistically significant improvements of model A over model B at the 95% confidence level are displayed in red, and statistically significant deterioration is in dark blue.

Charron et al. (2010) stress that from a practical perspective the maintenance of several state-of-the-art subgrid parameterizations is challenging. However, when using a single set of subgrid parameterizations, their ensemble prediction system, even when including two different stochastic parameterization schemes, was not as skillful as their multiphysics ensemble. Our results are promising, in that the inclusion of a stochastic kinetic-energy

TABLE 3. Pair-wise comparison of the percentage of outcomes, where model A (columns) performs better or worse than model B (rows) as measured by the Brier score when verified against observations. The outcomes comprise the forecast lead times 12 and 60 h, 4 verification events (see text), and 7 vertical levels for the variables zonal wind u, meridional wind v, and temperature T, totaling 168 outcomes. The bold numbers in parentheses denote statistically significant outcomes at the 95% confidence level. The mean monthly bias was removed from each ensemble member prior to the verification.

	PHYS better	PHYS worse	STOCH better	STOCH worse	PHYS_STOCH better	PHYS_STOCH worse
CNTL	82 (**39**)	18 (**2**)	93 (**57**)	7 (**1**)	87 (**54**)	13 (**3**)
PHYS			63 (**14**)	37 (**5**)	79 (**31**)	21 (**3**)
STOCH					58 (**14**)	42 (**8**)

TABLE 4. As in Table 3, but for the verification against analyses. The outcomes comprise the forecast lead times 12 and 60 h, 4 verification events, and 32 vertical levels for the variables zonal wind u, meridional wind v, and temperature T totaling 768 outcomes. The bold numbers in parentheses denote statistically significant outcomes at the 95% confidence level.

	PHYS better	PHYS worse	STOCH better	STOCH worse
CNTL	65 (**37**)	35 (**7**)	86 (**71**)	14 (**5**)
PHYS			76 (**59**)	24 (**5**)

backscatter scheme could yield similar, if not better probabilistic skill than the multiphysics ensemble, making stochastic parameterizations a real alternative to "ensembles of opportunity."

An additional outcome of the present work is that—consistent with the findings of Palmer et al. (2009), Charron et al. (2010), and Hacker et al. (2011)—combining multiple stochastic parameterizations or stochastic parameterization with multiple physics suites resulted in the most skillful ensemble prediction system. There is no doubt that different model-error strategies represent fundamentally different forms of model error. Thus, a combination of multiple model-error representations seems best suited to capture the complex nature of model error and yield the most reliable ensemble system.

Acknowledgments. This work was partially supported by the Air Force Weather Agency. Thanks go to Matt Pocernich, who developed the verification package that was used to produce Figs. 12–16. We are indebted to David Gill, Julie Schramm, and Jimy Dudhia for their insight into the physical and technical aspects of WRF and the AFWA mesoscale ensemble. Thanks to Steven Rugg for his enthusiasm for stochastic parameterizations and to Steve Mullen for insightful comments on an earlier version of this manuscript.

REFERENCES

Berner, J., 2005: Linking nonlinearity and non-Gaussianity of planetary wave behavior by the Fokker–Planck equation. *J. Atmos. Sci.,* **62,** 2098–2117.

——, F. J. Doblas-Reyes, T. N. Palmer, G. Shutts, and A. Weisheimer, 2008: Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal predicition skill of a global climate model. *Philos. Trans. Roy. Soc. London,* **366A,** 2561–2579.

——, M. L. G. Shutts, and T. Palmer, 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.

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

——, ——, S. E. Beare, K. R. Mylne, and G. J. Shutts, 2009: The local ETKF and SKEB: Upgrades to the MOGREPS short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **135,** 767–776.

Buizza, R., 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.

——, P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP Global Ensemble Prediction Systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Charron, M., G. Pellerin, L. Spacek, P. L. Houtekamer, N. Gagnon, H. L. Mitchell, and L. Michelin, 2010: Toward random sampling of model error in the Canadian Ensemble Prediction System. *Mon. Wea. Rev.,* **138,** 1877–1901.

Doblas-Reyes, F., and Coauthors, 2009: Addressing model uncertainty in seasonal and annual dynamical seasonal forecasts. *Quart. J. Roy. Meteor. Soc.,* **135,** 1538–1559.

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

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Frederiksen, J. S., and A. G. Davies, 1997: Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *J. Atmos. Sci.,* **54,** 2475–2492.

——, and ——, 2004: The regularized DIA closure for two-dimensional turbulence. *Geophys. Astrophys. Fluid Dyn.,* **98,** 203–223.

——, and S. M. Kepert, 2006: Dynamical subgrid-scale parameterizations from direct numerical simulations. *J. Atmos. Sci.,* **63,** 3006–3019.

Hacker, J. P., and Coauthors, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: Scientific description and performance results. *Tellus,* in press, doi:10.1111/j.1600-0870.2010.00497.x.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus,* **57A,** 219–233.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioners Guide in Atmospheric Science*. Wiley and Sons, 240 pp.

Kalnay, E., M. Kanamitsu, and W. Baker, 1990: Global numerical weather prediction at the National Meteorological Center. *Bull. Amer. Meteor. Soc.,* **71,** 1410–1428.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Li, X., M. Charron, L. Spacek, and G. Candille, 2008: A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Wea. Rev.,* **136,** 443–462.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Mason, P., and D. Thomson, 1992: Stochastic backscatter in large-eddy simulations of boundary layers. *J. Fluid Mech.,* **242,** 51–78.

Molteni, F., and T. N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. *Quart. J. Roy. Meteor. Soc.,* **119,** 269–298.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 2004: Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature,* **430,** 768–772.

Palmer, T. N., 2001: A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parameterization in weather and climate prediction. *Quart. J. Roy. Meteor. Soc.,* **127,** 279–304.

——, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 44 pp. [Available online at http://www.ecmwf.int/publications/library/ecpublications/_pdf/tm/501-600/tm598.pdf.]

Penland, C., 2003: Noise out of chaos and why it won't go away. *Bull. Amer. Meteor. Soc.,* **84,** 921–925.

Plant, R. S., and G. C. Craig, 2008: A stochastic parameterization for deep convection based on equilibrium statistics. *J. Atmos. Sci.,* **65,** 87–105.

Santanello, J. A., Jr., C. D. Peters-Lidard, S. V. Kumar, C. Alonge, and W.-K. Tao, 2009: A modeling and observational framework for diagnosing local land–atmosphere coupling on diurnal time scales. *J. Hydrometeor.,* **10,** 577–599.

Sardeshmukh, P., C. Penland, and M. Newman, 2001: Rossby waves in a fluctuating medium. *Progress in Probability: Stochastic Climate Models,* P. Imkeller and J.-S. von Storch, Eds., Vol. 49, Birkäuser Verlag, 369–384.

Shutts, G. J., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.,* **131,** 3079–3102.

——, and T. N. Palmer, 2007: Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate,* **20,** 187–202.

Skamarock, W. C., and Coauthors, 2008: A description of the Advanced Research WRF Version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp.

Stainforth, D., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature,* **433,** 403–406.

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

Teixeira, L., and C. A. Reynolds, 2008: Stochastic nature of physical parameterizations in ensemble prediction: A stochastic convection approach. *Mon. Wea. Rev.,* **136,** 483–496.

Tennant, W., G. Shutts, and A. Arribas, 2011: Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Wea. Rev.,* **139,** 1190–1206.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

von Storch, H., and F. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 494 pp.

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

Wilks, D., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 464 pp.