2011

Linear and non-linear response to parameter variations in a mesoscale model

Hacker, J.P.

þÿ T e l l u s (2 0 1 1) , 6 3 A , p p . 4 2 9 4 4 4

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

Tellus (2011), 63A, 429–444
Printed in Singapore. All rights reserved

TELLUS

Linear and non-linear response to parameter variations in a mesoscale model

By J. P. HACKER[1]*, C. SNYDER[2], S.-Y. HA[2] and M. POCERNICH[2], [1]*Naval Postgraduate School, Department of Meteorology, 589 Dyer Rd., Monterey, CA, USA;*[2]*National Center for Atmospheric Research, Boulder, CO, USA*

(Manuscript received 25 May 2010; in final form 15 December 2010)

ABSTRACT

Parameter uncertainty in atmospheric model forcing and closure schemes has motivated both parameter estimation with data assimilation and use of pre-specified distributions to simulate model uncertainty in short-range ensemble prediction. This work assesses the potential for parameter estimation and ensemble prediction by analysing 2 months of mesoscale ensemble predictions in which each member uses distinct, and fixed, settings for four model parameters. A space-filling parameter selection design leads to a unique parameter set for each ensemble member. An experiment to test linear scaling between parameter distribution width and ensemble spread shows the lack of a general linear response to parameters. Individual member near-surface spatial means, spatial variances and skill show that perturbed models are typically indistinguishable. Parameter–state rank correlation fields are not statistically significant, although the presence of other sources of noise may mask true correlations. Results suggest that ensemble prediction using perturbed parameters may be a simple complement to more complex model-error simulation methods, but that parameter estimation may prove difficult or costly for real mesoscale numerical weather prediction applications.

1. Introduction

A canonical approach to accounting for uncertainty in mesoscale numerical weather prediction (NWP) model formulation is to use several different models in an ensemble prediction system (e.g. 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; Hacker et al., 2011). Models are different if they contain different equation sets, where the differences may be in the resolved-scale dynamics or the subgrid scale and forcing schemes (so-called model 'physics'). Published literature typically distinguishes between ensembles using only different physics (multiphysics or multischeme) and ensembles that include different dynamical equations (multimodel). Empirical evidence for predictive probabilistic skill from the use of multimodel or multiphysics ensembles has provided motivation for continuing to deploy and study them.

Multimodel or multiphysics ensembles have some drawbacks. Because each member of the ensemble is a realization from a different distribution, and will predict a trajectory on a different attractor, it is not obvious how to interpret ensemble covariances. Thus a powerful outcome of an ensemble prediction may be compromised. Secondly, biases of individual members also need to be considered, and possibly removed (e.g. Eckel and Mass, 2005) before forming probabilistic predictions. Although methods exist for calibrating multimodel ensembles, such as the standard inference technique of Bayesian model averaging (Raftery et al., 2005), different error distributions for each member complicates calibration efforts. Thirdly and more practically, multimodel or multiphysics ensembles incur greater development and maintenance costs (as pointed out by Bowler et al., 2008). For example, a 10 member multiphysics ensemble with unique microphysics, turbulence, radiation and deep convection schemes for each member would require a total of 40 schemes. Development of each is typically a multiyear effort. While this may be feasible for community models, operational forecast centres must still subject each scheme or set of schemes (a suite) to thorough testing. Finally, changing schemes may necessitate changing the variables carried by the model; for example, different microphysics schemes represent different species of water, and some combine multiple hydrometeors into a single variable. Thus, identical use across members of model output in order to drive secondary or diagnostic models becomes impossible. Difficulties with interpreting, developing and maintaining them motivate efforts to find alternatives.

*Corresponding author.
e-mail: jphacker@nps.edu
DOI: 10.1111/j.1600-0870.2010.00505.x

Perturbing parameters within a single physics suite offers an alternative to multimodel ensembles. It eliminates the need to develop and maintain multiple schemes, and forecast centres can devote resources to finding optimal sets of parameters. Forecasts using each parameter set would still need to be evaluated to avoid poor performance, and the performance response to parameter variations may yield information helpful for improving the schemes themselves. The drawback to a perturbed-parameter approach is that it cannot represent all model errors, such as when model structure is incorrect in addition to errors in parameter specification.

Perturbed parameters have been used recently in ensemble-prediction studies, with some success. Murphy et al. (2004) and Stainforth et al. (2005) describe climate-prediction ensembles using perturbed parameters, finding sensitivity and rejecting some parameter sets based on poor fits to past climate. Bowler et al. (2008) described the use of multiple parameters, each evolving in time with an auto-regressive process to add stochasticity, in the Met Office ensemble prediction system. Although it helped the ensemble performance, a study of systematic or linear responses was not reported. Detailed characterization of the mesoscale response to the perturbations at weather-prediction time scales is still lacking in the literature.

Given observations and an ensemble making use of perturbed parameters, one possible approach to improving the model is estimation of parameters and their uncertainty through data assimilation. Theory for parameter estimation, particularly with linear state estimation techniques, is well established (cf. Cohn, 1997). Annan et al. (2005a,b) demonstrated the potential for estimation of several parameters in an ocean model and atmospheric GCM, respectively. Aksoy et al. (2006a,b) examined the potential in a 2-D sea breeze model and a mesoscale NWP model, respectively. More recently, Tong and Xue (2008) examined the potential for estimation of microphysical parameters with an ensemble filter, finding that estimating multiple parameters is difficult because the strength of the parameter–state relationship decreases as the number of parameters increased. Posselt and Vukećević (2010) also considered microphysical parameters but emphasized the complex relationship between those parameters and the microphysical fields, without attempting parameter estimation.

Successful parameter estimation requires a robust relationship between parameters and state variables that are observable, and a linear or at least monotonic relationship eases the problem considerably. Nielsen-Gammon et al. (2010) refers to a parameter that results in a reasonably strong and unique relationship between parameters and state variables as distinguishable, and we adopt the same terminology.

In this work, we explore relationships between parameters and predictions to address some general outstanding questions related to varying parameters in a realistic mesoscale ensemble forecast system. Defining 'reasonable' parameter variations to be near the range of expected uncertainty found in the literature, the goals can be summarized broadly as follows.

(i) To quantify the ensemble spread in near-surface predictions produced by reasonable parameter variations.

(ii) To determine whether reasonable parameter variations necessarily create an inferior or superior model.

(iii) To find relationships between parameters and predictions that can be exploited with linear data assimilation techniques.

Unlike most previous studies, we consider variations of multiple parameters simultaneously and within the context of a full ensemble-forecasting system, in which the ensemble is also influenced by uncertain initial conditions (ICs) and boundary conditions. Our results concerning relationships between parameters and predictions, or the lack thereof, are necessarily more pessimistic than previous studies; that is, we find little evidence that strong linear parameter–state relationships are common. Nevertheless, to the extent that model solutions may depend nonlinearly on the parameter variations, experiments with multiple parameters and with other sources of noise (such as initial and boundary conditions) are necessary to evaluate how parameter variations will influence forecasts in practice.

Most of our experiments are also restricted to small ensembles (10 members) owing to our interest in specific operational applications (Hacker et al., 2011), which include computational limitations. With such small ensembles we are less likely to accurately estimate correlations, but strong correlations should still be measurable. We apply field significance tests to assess, for each forecast and each lead time, whether the parameter–state correlations computed from the ensemble are a chance occurrence or indicative of a robust parameter–state relationship that could be the basis for parameter estimation within a wide range of flows. Results show that the forecasts are dominated by lack of significant correlation, but intermittent and spatially coherent patterns of significant correlations can exist.

Section 2 briefly explains the ensemble design, parameters explored and methods for choosing them. Section 3 presents a demonstration of forecast sensitivity to individual parameters, and provides an example of the ensemble spread resulting from simultaneous parameter perturbations. Those results provide context for Section 4, which explores distinguishability resulting from parameter perturbations with individual-member forecast means and errors. Section 5 presents spatial structures of correlations between parameters and forecasts, including field significance testing, to understand the frequency of robust correlations. Further comments and summary are in Section 6.

2. Ensemble design

Ensemble-member diversity in this experiment comes from three sources intended to capture some part of ICs, boundary-condition and model uncertainty. A global ensemble provides ICs and lateral boundary conditions (LBCs) for forecasts with

the Advanced Research version of the Weather Research and Forecast (WRF) model (Skamarock et al., 2008). Multiple parameter sets within a single physics suite accounts for some model uncertainty within the mesoscale WRF domain ($\Delta X = 45$ km). Except where noted, the ensemble contains 10 members.

The global ensemble is the U.S. National Centers for Environmental Prediction (NCEP) global ensemble forecast system (GEFS; see Wei et al., 2008). In GEFS, an ensemble transform rescales perturbations every 6 h, and recentres them on NCEP's operational deterministic analysis given by the NCEP Gridpoint Statistical Interpolation [GSI; Kleist et al. (2009)] 3-D variational data assimilation scheme. A new set of perturbations are computed every 6 h, and each is rescaled according to the analysis error variance as specified in the GSI. The ensemble is delivered with a horizontal grid spacing of 1°, a vertical grid spacing of 25 hPa and a time step of 6 h. We use the first 10 members for the WRF ensemble.

Sea surface temperature (SST) analyses from the U.S. Navy, and soil analyses from the U.S. Air Force, provide surface and subsurface initialization. Each ensemble member uses the same static SST field for an individual forecast; soil conditions are initially identical but evolve in the WRF integrations. Land-surface uncertainty is included via perturbations imposed to individual land-cover categories, as described in detail in Hacker et al. (2011). Here the land-surface perturbations are simply additional sources of noise in the ensemble.

We next describe model perturbations constructed by altering a few parameters in the control physics suite, which is that of member six in table 1 of Hacker et al. (2011). One cannot hope to explore perturbing all uncertain parameters in the model, especially in this initial study. The choices here resulted from conversations with model and parametrization developers, and we cannot claim to have chosen the parameters eliciting the most sensitivity possible.

Perturbing many parameters need not lead to qualitatively different or even larger variations in the model solutions, since multiple parameter settings may produce the same output from a physical parametrization. Posselt and Vukećević (2010) found this behaviour when microphysical parameters were varied in single-column cloud model, and Alapaty et al. (1997) found the same with parameters in a boundary-layer (PBL) model. We therefore choose to perturb only a single parameter in each of four physical parametrizations: those for cumulus convection, the PBL, microphysics and short-wave radiation.

2.1. Description of parameters

Table 1 summarizes the set of parameters perturbed for this experiment and the distributions initially specified for each. The Kain–Fritsch (KF) scheme is a mass-flux cumulus parametrization that contains an entraining and detraining plume model. Its origins are documented in Fritsch and Chappel (1980), Kain and Fritsch (1990) described modifications to the plume model, and Kain (2004) describes further modifications. Its solution is sensitive to the parameter describing the subgrid-cloud radius (R), which controls the maximum possible entrainment rate in the plume model. Originally, R was specified as a constant, but more recent versions consider it a diagnostic variable. Kain (2004) states that 'we have little or no skill in actually predicting what the horizontal dimensions of convective clouds in the atmosphere will be'.

Currently, R is parametrized as a function of $W_{\mathrm{KL}} = w_{\mathrm{g}} - c(Z_{\mathrm{LCL}})$, where w_{g} is an approximate resolved vertical velocity near the lifting condensation level (LCL) and $c(Z_{\mathrm{LCL}})$ is a threshold vertical velocity that depends on the height of the LCL. W_{KL} is used in computing a temperature perturbation for the trigger function in the KF scheme. R can take a value of 1000 ($W_{\mathrm{KL}} < 1000$), $1000 + W_{\mathrm{KL}}/10$ ($1000 \leq W_{\mathrm{KL}} \leq 2000$) or 2000 m ($W_{\mathrm{KL}} > 2000$). We represent uncertainty in R with an additive perturbation drawn from the distribution described in Table 1. This perturbation is fundamentally different from the others because it is added onto a time-dependent diagnostic variable.

The Yonsei University (YSU) boundary-layer scheme is an extension of the scheme first developed by Troen and Mahrt (1986), based on K-theory with a counter-gradient mixing term and later updated by Hong and Pan (1996). Noh et al. (2003) modified the mixing profile to include explicit treatment of the entrainment rate at the top of the mixed PBL. The entrainment rate is a function of a coefficient A_R and a velocity scale, where the functional form of the velocity scale changes between states of free-convection and the presence of shear. Noh et al. (2003) recommends a value of $A_R = 0.15$ based on LES experiments, agreeing with Moeng and Sullivan (1994). From qualitative

Table 1. Parameters or variables chosen for the perturbation experiments, with descriptions of the initial distributions assigned

Scheme	Parameter	Units	Min	Mean	Max	Distribution
KF	ΔR	m	-300	0	300	$\beta(6, 6)$
YSU	A_R	None	0.1	0.15	0.3	$\beta(2, 6)$
WSM5	N_0	m^{-4}	2×10^6	8×10^6	2×10^9	$\beta(1.5, 6)$
Dudhia	α_{CA}	$\mathrm{m}^2\ \mathrm{kg}^{-1}$	2×10^{-6}	1×10^{-5}	2×10^{-5}	$\beta(4.8, 6)$

Note: See text for a description of each parameter.

arguments, Ball (1960) suggests $0.1 \leq A_R \leq 0.3$, which we adopt to define the distribution for drawing perturbations.

Hong et al. (2004) details the WRF Single-Moment Five-class (WSM5) microphysics scheme. Water vapour, rain, snow, cloud ice and cloud liquid water are handled separately, and supercooled water is permitted. Properties of these single-moment schemes are sensitive to specifications of rain-drop and ice-particle size distributions. Although it may elicit less sensitivity during the winter over the continental United State, we choose to perturb the intercept parameter in the exponential rain drop-size distribution. The intercept parameter (N_0) is almost universally taken to be 8×10^6 m^{-4}, following the results of Marshall and Palmer (1948). But observational studies show that N_0 can vary by at least an order of magnitude (e.g. Waldvogel, 1974; Sauvageot and Lacaux, 1995).

Dudhia (1989) describes the short-wave radiative transfer model used here. It is a simple downward integration of solar flux, accounting for water vapour and cloud absorption, cloud albedo and clear-air scattering. The percentage of solar irradiance scattered in a model layer is directly proportional to the layer-integrated density of the dry air and the scattering parameter α_{CA}. The WRF default value is 1×10^{-5} m^2 kg^{-1}, but in reality depends on the local composition of atmospheric constituents.

First-order sensitivity to each parameter can be predicted from the parametrization equations as follows.

(i) Greater low-level vertical velocity in the resolved dynamics leads to a greater R in the KF scheme, and thus a lower value of maximum entrainment rate. R is inversely related to updraft dilution, and affects the vertical redistribution of heat, moisture and momentum in the closure.

(ii) PBL entrainment rate is directly proportional to A_R in the YSU scheme. Greater entrainment will promote PBL growth, and warm and dry the well-mixed PBL as free-tropospheric air is mixed downward.

(iii) The intercept parameter for rain in the WSM5 (as is typical for microphysics) directly influences the entire drop-size distribution for a given rain water content. The slope of the distribution is proportional to the intercept; decreasing N_0 shifts the mean concentration toward larger drops, and increasing N_0 shifts it toward smaller drops. To first-order, the rain rate will also shift proportionally to N_0.

(iv) The effects of more or less clear-air scattering in the Dudhia short-wave scheme are obvious. Less scattering leads to more direct solar irradiance at the tops of clouds and the surface, and vice versa.

Although the first-order effects of each of the parameter perturbations can be predicted, it is not clear that these approaches will lead to biases in the model solutions. Linear compensating effects and non-linear processes may play a role in propagating the parameter perturbations through the model state in complex ways that are difficult to predict. The analysis in Sections 4 and 5 addresses this issue.

In a small set of sensitivity experiments (not shown), we ran the WRF with single parameters set to either the minimum or maximum value shown in Table 1. Difference maps at multiple forecast lead times provide a measure of the maximum sensitivity possible from single-parameter perturbations within the specified distributions. Looking primarily at surface variables (2-m temperature and water vapour, 10-m winds), the magnitudes in the difference fields did not depend strongly on the parameter varied. This suggests the potential for similar contributions to ensemble spread from each parameter. Conversely, details in the difference field were sensitive to the parameter chosen, suggesting that each parameter can affect ensemble spread independently.

2.2. Parameter selection

The parameters form a vector (ΔR, A_{R}, N_0, α_{CA}), spanning the parameter space for this experiment. Lacking a priori knowledge of the broader effects of each parameter on the forecasts, our goal is to fill this parameter space with points that are nearly equally distant from each other.

One technique for this space-filling problem is modified Latin Hypercube Sampling (LHS). The experiment design is based on the hypothesis that the model solution resulting from any possible set of parameter values within the assigned limits is equally likely. To select the sets of parameters we assume: (1) each parameter can take any value within the specified distribution and (2) each parameter is independent from the others. A reasonable approach within this framework is to spread the parameter vectors evenly throughout the parameter space. LHS, with the additional constraint that the mean distance between points is nearly maximum, achieves this.

We refer the interested reader to Santer and Williams (2003) for details on maximum distance LHS, and here simply state that it satisfies our goal. LHS works directly in a normalized space so that draws in a four-dimensional space are made with each dimension as $\mathcal{U}(0.5, 1)$, and then the logistic transformation puts each draw into the appropriate β distribution. We verified our implementation by selecting 1000 sets and plotting the resulting distributions, successfully recovering the assigned β and confirming that the draw of each parameter is independent. Scatter plots between all possible pairs of parameters in 10 draws of the parameter vector (Fig. 1) show that the space is filled and the parameters are independent from one another.

After transforming the points in Fig. 1 to sets of parameter values, we ran a preliminary experiment with one month of ensemble forecasts in an east-Asian domain. Each set was assigned to a single ensemble member, and held constant throughout the experiment. We have no reason to believe that these values should be constant in time, but the appropriate temporal evolution is unknown.

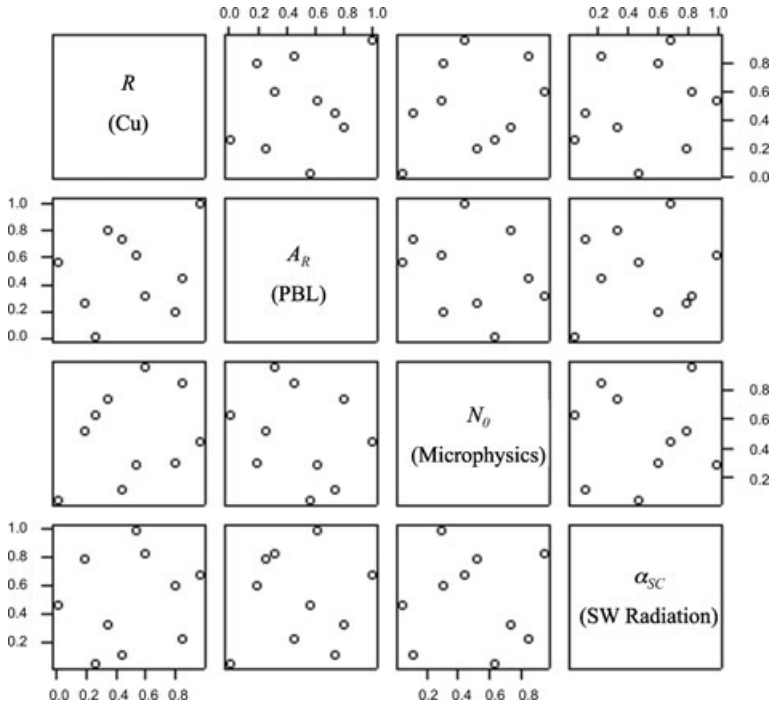

Fig. 1. Joint distributions for pairs of random-uniform draws using Latin Hypercube Sampling. These values are transformed to the distributions described in Table 1.

Examination of ensemble spread suggested limited impact from the parameters, compared to the spread introduced from initial and LBCs. Further metrics suggested nearly undetectable systematic relationships between parameter values and forecasts. As described next, broadening the parameter distributions provides both a test for linearity, and also new parameter sets used in the analysis.

2.3. Parameter scaling tests for linearity

Ensemble spread will scale with parameter-perturbation magnitude if linear responses are dominant. Analysis of preliminary experiments (not shown) did not reveal systematically monotonic relationships between parameters and predictions. But it is possible that the sets of perturbations drawn from the distributions in Table 1 were too small to easily extract a linear signal when the ensemble is subject to other sources that force variability.

The experiment presented in this section uses ensemble spread to test for linearity over a wider range of parameter values. Results shown are from a month-long (28 forecasts) experiment on a domain centred over the Korean Peninsula but with the same model configurations; the results of this test should not change qualitatively with the location or test period.

To choose new parameter ranges, we scale the parameter distributions by comparing the spreads of the multiparameter ensemble and a multiphysics ensemble (ensemble Phys in Hacker et al., 2011). Spatially averaged ensemble spread (variance) is dominated by large-scales, which are primarily controlled by the GEFS ensemble used for LBCs. We can assume spread from the different sources is additive, and make use of a single-model ensemble that directly downscales the GEFS ensemble (*Cntl* in Hacker et al., 2011). The ratios of spread in addition to *Cntl* is

$$\alpha^2 = \frac{\sigma^2_{Phys} - \sigma^2_{Cntl}}{\sigma^2_{Param} - \sigma^2_{Cntl}}, \tag{1}$$

where *Phys* denotes the multiphysics ensemble and Param denotes the multiparameter ensemble. Values of σ^2 for each lead time are computed as gridpoint ensemble variance averaged over the domain and forecasts. Before computing spreads, experiment-mean forecast fields for each forecast lead time and member are subtracted so that different model biases do not contribute to spread.

Values of α are shown in Fig. 2 . The average of α^2 from 6–48 h leads to $\alpha \approx 7.5$, which is used to scale the parameter distributions. For simplicity, we treat the distributions as if they are Gaussian (logarithms are used for N_0 to make it more Gaussian-like), and scale the variance of the parameters. Because it is impossible to scale the variances by a factor of 7.5^2 with only 10 members and the bounds in Table 1, parameter bounds for ΔR, A_R and α_{CA} are adjusted to the values in Table 2. The results from scaling, compared to the original parameter values, are shown in Fig. 3. Rerunning the experiment with the expanded parameter ranges does not compromise model stability.

Spreads after subtracting σ^2_{Cntl}, computed from ensembles run with the new parameter values, show that the spread does not scale with the spread in the parameter distributions (Fig. 4). For comparison, the spread predicted by the linear assumption and

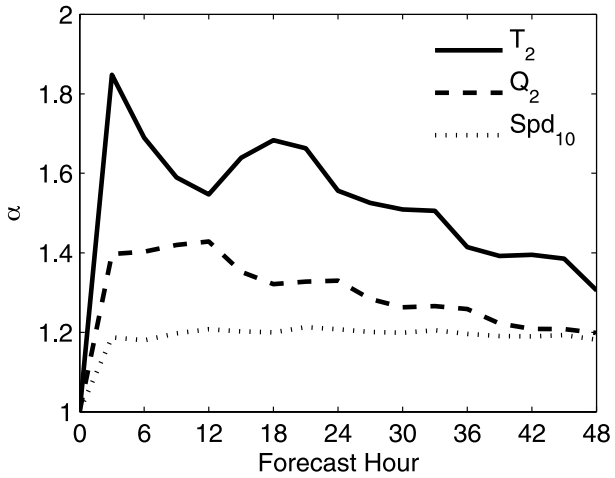

Fig. 2. Square-root of the ratios of multiphysics ensemble spread to multiparameter ensemble spreads after subtracting the spread of the single-model ensemble *Cntl* (eq. 1).

Table 2. Expanded parameter bounds, needed to scale up the parameter distribution variance according to the experiments in Section 2.3

Scheme	Parameter	Units	Min	Max
KF	ΔR	m	−999	999
YSU	A_R	None	0.05	0.35
WSM5	N_0	m^{-4}	2^6	2×10^9
Dudhia	α_{CA}	$m^2\ kg^{-1}$	2×10^{-7}	2.5×10^{-5}

Note: Minima and maxima can be compared to those in Table 1, and are used in the remaining analysis.

α is also shown in Fig. 4 with a thin dashed curve. Spreads increase with the parameter scaling most of the time, but do not approach predicted values.

The scaling test performed here shows that a wider distribution of parameters can increase the spread, but spread does not increase according to a linear relationship between the parameters and the forecasts. We conclude that a linear response is generally not evident over a broad parameter range, and further investigate distinguishability and monotonicity in Sections 4 and 5.

3. Sensitivity to parameter perturbations

A probabilistic verification of this ensemble is presented in Hacker et al. (2011), where is it called Param. It is shown to be less skilful than an ensemble using multiple physical parametrization schemes, but more skilful than an ensemble with no variations in the mesoscale model (i.e. a direct dynamical downscaling of the GEFS ensemble, called *Cntl* in Hacker et al., 2011). For the purposes of the present work, it is sufficient to know that parameter perturbations do not produce an ensemble inferior to *Cntl*. To gain further intuition about the effects of parameter perturbations, we examine forecast differences from perturbing the individual parameters, evaluate the ensemble spread introduced from parameter distributions relative to *Cntl*, and provide an example of that spread.

We can quantify the effect of perturbing individual parameters with either the minimum or maximum value of a single parameter (Table 2) to provide context for runs with multiple parameter perturbations. 12 forecast periods, each initialized at 00 UTC and separated by 5 d, provide a range of synoptic conditions while keeping computational requirements to a minimum. Meridional wind anomalies suggest baroclinic systems over the continental United States during 6 of the 12 forecasts. One of each of the four parameters is given its minimum or maximum value, for a total of eight model perturbations applied to each forecast period. All of the forecasts in these 'ensembles' use the same initial and LBCs, thereby isolating the effect of an individual parameter. Tests are on the the domain analysed in the remainder of this paper, over the continental United States, providing information somewhat independent from the scaling tests over Asia.

Mean absolute differences (MAD) between two forecasts with perturbations to an individual parameter show that although important differences can be identified, each parameter is capable of producing a response that is usually the same order of magnitude as the other parameters (Fig. 5). The parameter α_{SC}, which is inversely proportional to solar insolation at the surface, produces a diurnal cycle most obvious for 2-m temperature; less diurnal variability is apparent for 10-m wind speed and 2-m water–vapour mixing ratio. Note here, that the wind speed MAD is not the length of the vector difference, which would be greater. The 00 UTC initialization time delays most of the effects until 12-h later, the following morning. Parameter A_R, controlling entrainment in the convective PBL, produces a smaller-amplitude diurnal variation in the MAD. Compared to α_{SC}, A_R perturbations produce differences in near-surface conditions that appear to persist more through the night. At the 00 UTC initialization time the PBL can still be convective over parts of the United States (e.g. the Southwestern deserts), and the effects can be felt immediately. Perturbations to parameters ΔR, A_R and N_0 produce similar-magnitude responses, except from precipitation accumulated during the forecast. Neither A_R nor α_{SC} perturbations contribute much to precipitation differences during the few remaining daylight hours immediately following initialization. Parameters that act directly on cloud processes (N_0 and R) act much more quickly during that period, and it is not until the following day that parameters A_R and α_{SC} produce noticeable precipitation differences.

Forecast spread resulting from simultaneously perturbing all four parameters can be evaluated with reference to the individual parameter results presented above. The model runs supporting the analysis above lack uncertainty from other sources, and the response to parameter perturbations in an ensemble system with

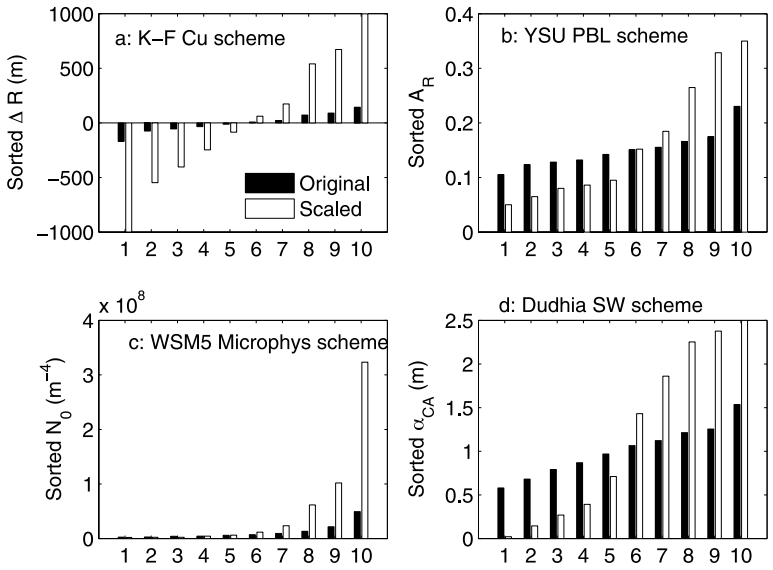

Fig. 3. Original parameter values corresponding to distributions specified in Table 1 (black) and parameter values after scaling the distributions by α (white). Values are sorted for presentation.

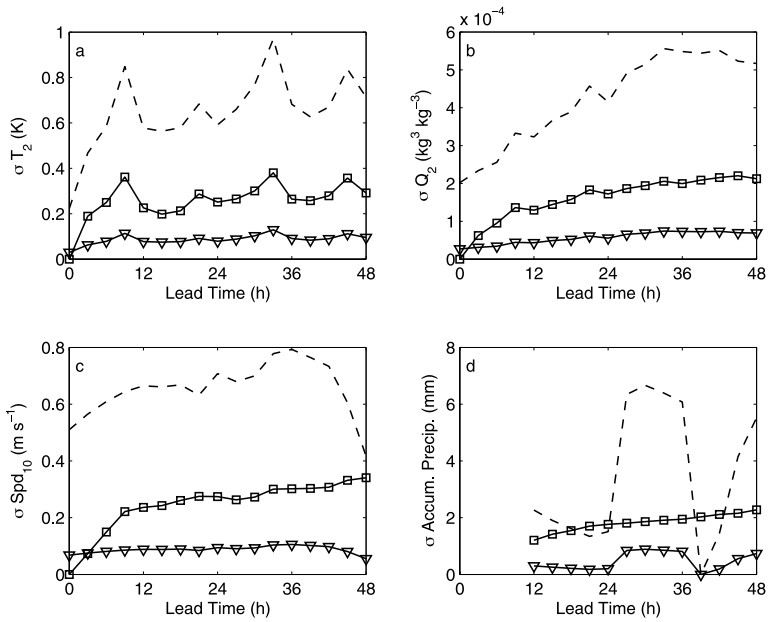

Fig. 4. Ensemble spreads after subtracting the spread of the downscaled ensemble *Cntl* (no model perturbations). Plotted are spreads resulting from the original parameter distributions (inverted triangle), spreads predicted by applying the scaling parameter α to the original spread (thin dashed) and spreads resulting from running the experiment with the scaled parameter distributions (squares). Results are shown for (a) 2-m temperature, (b) 2-m water vapour mixing ratio, (c) 10-m winds and (d) accumulated precipitation.

additional sources of variability will be more difficult to detect. Because we are interested in the more typical case of an ensemble forecast system influenced by other perturbations, we consider the ensemble subject to large-scale initial and boundary condition variability from the global ensemble. The following analysis uses 64 ensemble forecasts and includes initialization at both 00 and 12 UTC.

Figure 6 shows that parameter perturbations increase spread compared to a single-model ensemble, with relative increases up to approximately 30% for 2-m temperature. Parameter perturbations induce a faster growth in spread during the first 12 h, and subsequent growth rates are similar. Additional spread from parameter perturbations is broadly similar in magnitude to the 12-h MAD seen in Fig. 5, except that 10-m wind speed is negligibly affected and the additional spread does not continue to grow throughout the forecast. The experiment-mean prediction for each ensemble member, lead time and gridpoint (i.e. the mean field) can be removed to simulate bias removal, giving the dashed curves in Fig. 6. Negligible differences between the curves with circles show that each of the *Cntl* ensemble members have similar bias, as expected. Land-surface property perturbations also have a negligible effect on domain-wide biases (not shown).

Greater differences between the curves with inverted triangles show that local bias removal reduces 2-m temperature spread by approximately 0.1 K, and water–vapour mixing ratio spread by

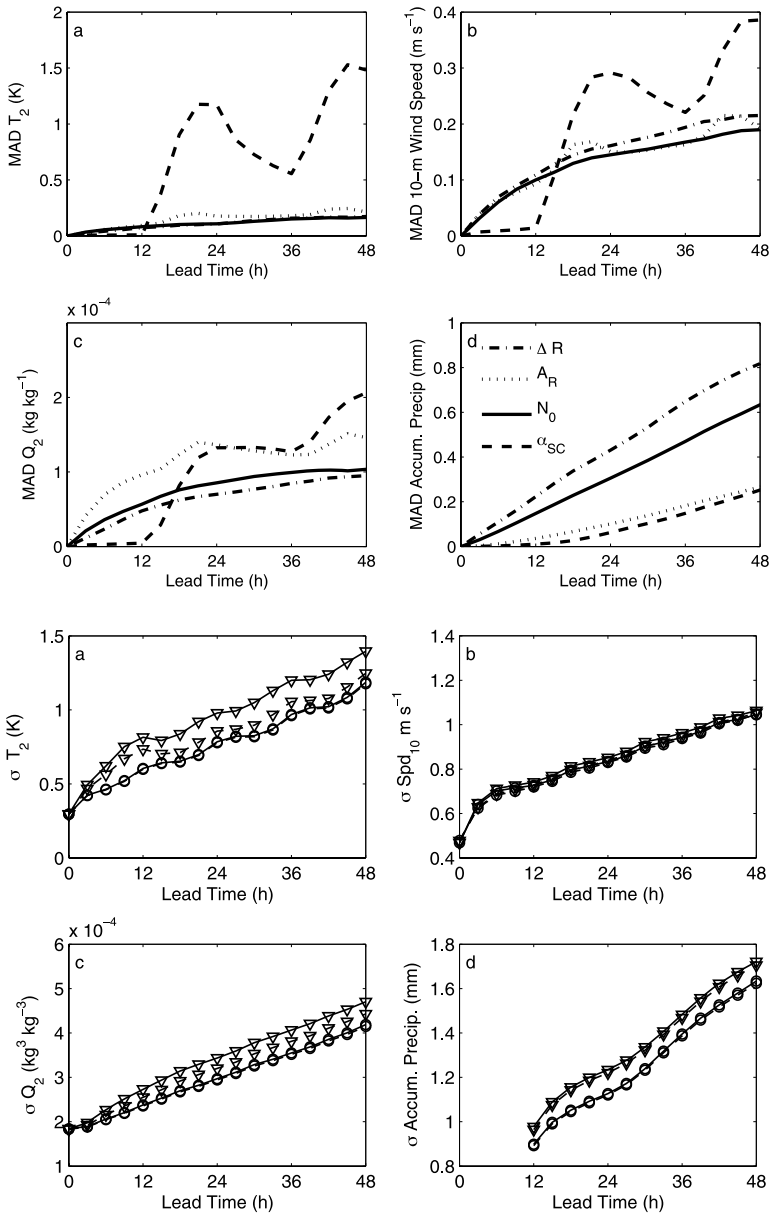

Fig. 5. Mean absolute difference (MAD) between forecasts resulting when a single parameter (noted in legend) is perturbed to either its minimum or maximum value (Table 2). The MADs for (a) 2-m temperature, (b) 2-m water vapour mixing ratio, (c) 10-m winds and (d) accumulated precipitation are computed from 12 forecasts. Each forecast is initialized at 00 UTC during November 2008–January 2009 over the continental United States, and separated by 5 d.

Fig. 6. Gridpoint ensemble spreads of predicted (a) 2-m temperature, (b) 2-m water–vapour mixing ratio, (c) 10-m wind speed and (d) accumulated precipitation. Results before (solid) and after (dashed) removal of the experiment-mean forecast field from each member. Inverted triangles denote the multiparameter ensemble and circles denote the single-model ensemble.

approximately 0.00001 kg^3 kg^{-3}. Spread in 10-m wind speed and accumulated precipitation is affected to a smaller degree. Removing biases removes more spread from the multiphysics ensemble than from the multiparameter ensemble (not shown).

An example of ensemble spread in 48-h predictions of 2-m temperature spread is shown in Fig. 7. Panel (a) shows results from the multiparameter ensemble and panel (b) shows results from the single-model (*Cntl*) ensemble, valid 0000 UTC 16 January 2009. The patterns are broadly similar, but details differ. Multiple parameters increase spread most notably over the Rocky Mountains and the west coast. More careful examination shows that increased spread from parameter perturbations occurs in the regions of greatest spread in *Cntl*. We next examine whether clear systematic forecast differences result from parameter perturbations.

4. Individual ensemble member distinguishability

In an ensemble forecast, indistinguishable members are indicative of equally plausible model implementations. This is consistent with the notion that a large number of possibly reasonable parameter sets exists, and that any one set is subject to uncertainty. Because the parameter sets do not change throughout the experiment, distinguishable members can be identified by looking at first and second-moment metrics. Predicted state mean

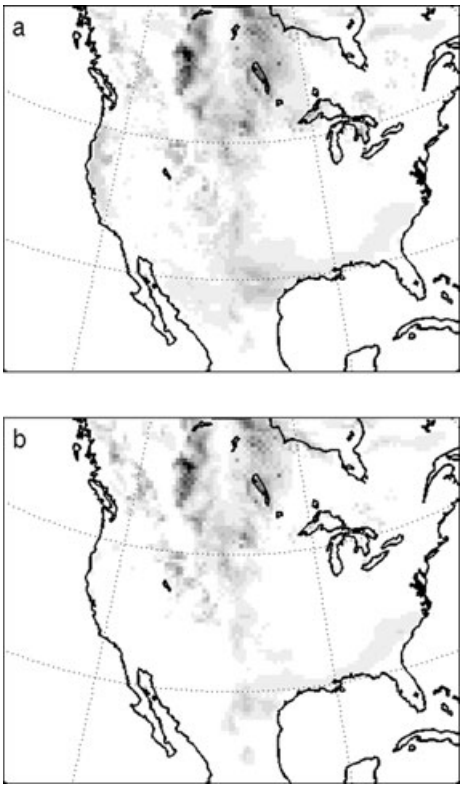

Fig. 7. Example 2-m temperature ensemble spread valid at 0000 UTC 16 January 2009 (48-h prediction) from the (a) multiparameter ensemble and (b) single-model ensemble. The grey scale corresponds to 0–35 K.

and variance, and error mean and variance. Here we show that with a few exceptions, individual members are not easily identified with domain-averaged quantities computed from surface forecasts.

4.1. Dependence of domain mean and variance on parameters

Consider 64-element samples formed from the domain (spatial) means or variances of all 64 forecasts by each individual ensemble member and lead time. Differing distributions indicate systematically distinguishable members. Further, if the differences appear to be a function of parameter values, then we might attribute those differences to the parameters themselves. Box plots (Figs 8–10) are useful to gain intuition about both the systematic behaviour and the variability introduced by LBCs, initialization and case-by-case variability in the weather.

All models are biased; different means indicate different biases, and biases near the surface can become apparent quickly. By showing results for 48-h lead times we ensure that the spatial means, and in particular the means of the distributions of the spatial means, are attributable to each ensemble member and not the ICs or any imbalances in initial states. During the research, we examined results from forecast lead times of 6–60 h, and surface variables 10-m winds, 2-m temperature and water–vapour mixing ratio, and accumulated precipitation. In nearly all instances, no relationship between spatial means and parameter values is evident. Here we show the exceptions.

Domain-mean forecasts with each member indicate that forecasts are generally not a monotonic function of parameter A_R, but one outlier in 10-m wind speed (Fig. 8b) is evident. We expect that a greater entrainment rate will mix more dry free-tropospheric air, with greater momentum, downward to increase near-surface winds and decrease humidity. Water–vapour mixing ratio (Fig. 8c) suggests a weak tendency to dry with greater A_R, but the signal may not be robust given the day-to-day variability among forecasts. Wind speed shows no sign of a trend, and instead the member perturbed with $A_R \approx 0.18$ can be characterized by systematically greater wind speed. This is ensemble Member 1, which is subject to small parameter perturbations (i.e. no perturbations near the distribution extremes).

For comparison, Fig. 9 shows spatial-mean predictions versus α_{SC}. The nearly monotonic temperature variation in Fig. 9a is the clearest indication we could find that individual parameter variations can be easily detectable. The median of the spatial-mean forecasts decreases by more than 1 K when clear-air scattering increases by two-orders of magnitude. The distributions show large overlap; we cannot say whether the relationship is robust although it appears possible. In Fig. 9b, the member with large wind speeds is again Member 1, and no other relationship with wind speed is apparent. Plots of spatial-mean 10-m wind speed against the other parameters (omitted) also reveal Member 1 as an outlier.

A parameter that changes the behaviour of a model could also change the spatial variance of predictions, but Fig. 10 shows that these parameter perturbations have no noticeable effect. This result may at first be expected because no stochasticity is directly introduced. But the parameters can possibly change forcing in the model and excite modes not present with the default parameter set. A simple example is the change in variance observed when changing the value of the constant forcing parameter F in the Lorenz (1963) three-variable model often used as a proxy for atmospheric dynamics.

Comparing spatial means or variances to parameter values cannot prove a causal relationship, but the lack of any systematic dependence of the predictions on the parameter values can reject one. We could not find robust evidence to suggest a relationship between parameter values and predictions. Viewing the predictions against a single parameter treats other sources of uncertainty as noise. Completing the same analysis for the land-use perturbations (not shown), we find again that Member 1 stands out in wind speed, but no other relationships are detectable.

We conclude that within the noise introduced via the GEFS, ICs and LBCs, and the weather itself, members are not

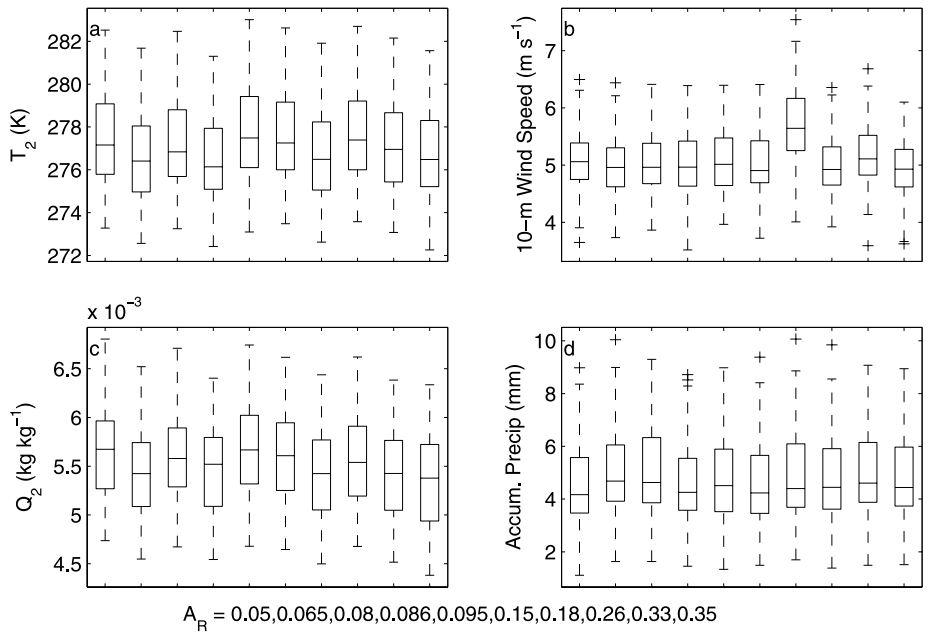

Fig. 8. Distributions of 48-h spatial-mean surface predictions versus A_R parameter values: (a) 2-m temperature, (b) 10-m wind speed, (c) 2-m water vapour mixing ratio and (d) accumulated precipitation. Boxes show lower quartile, median and upper quartile. Whiskers show 1.5 times the interquartile range, and + signs denote outliers. Parameter values are indicated in the text along the bottom; values are sorted for presentation.

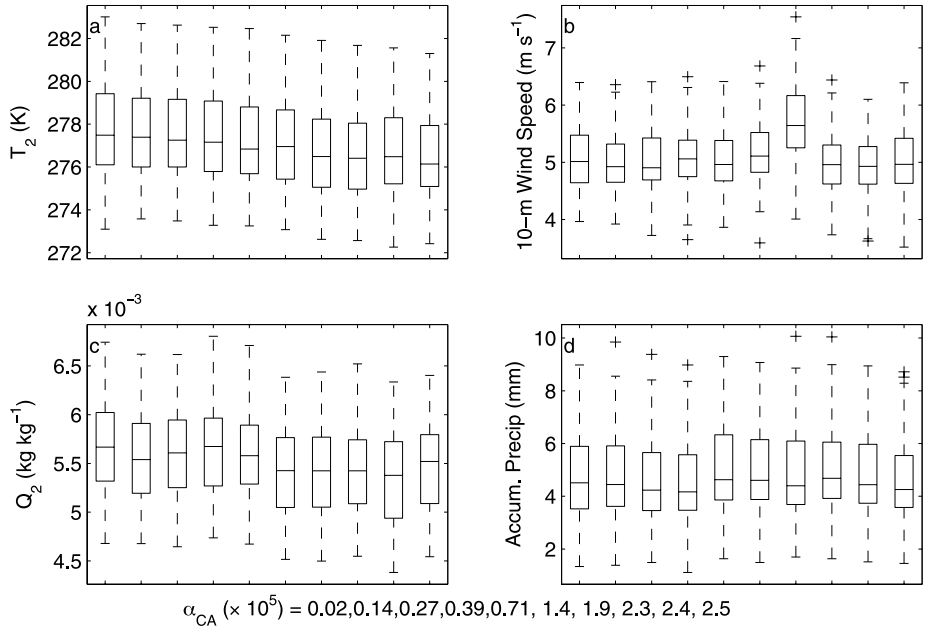

Fig. 9. Same as Fig. 8 but for $\alpha_{CA} \times 10^5$ parameter values.

readily distinguishable from domain-mean quantities. We note that other recent studies (e.g. Nielsen-Gammon et al., 2010) have taken a more local approach by dividing up a domain into regions a priori. One example is to take averages of only points over land or only over water. Instead our goal is to identify domain-wide systematic differences. We make no a priori assumptions about the spatial extent over which a parameter should be varied, and choose to keep the approach as general as possible.

4.2. *Individual-member near surface errors*

Somewhat related to the distributions of the states shown above, error (observation minus forecast) distributions can be

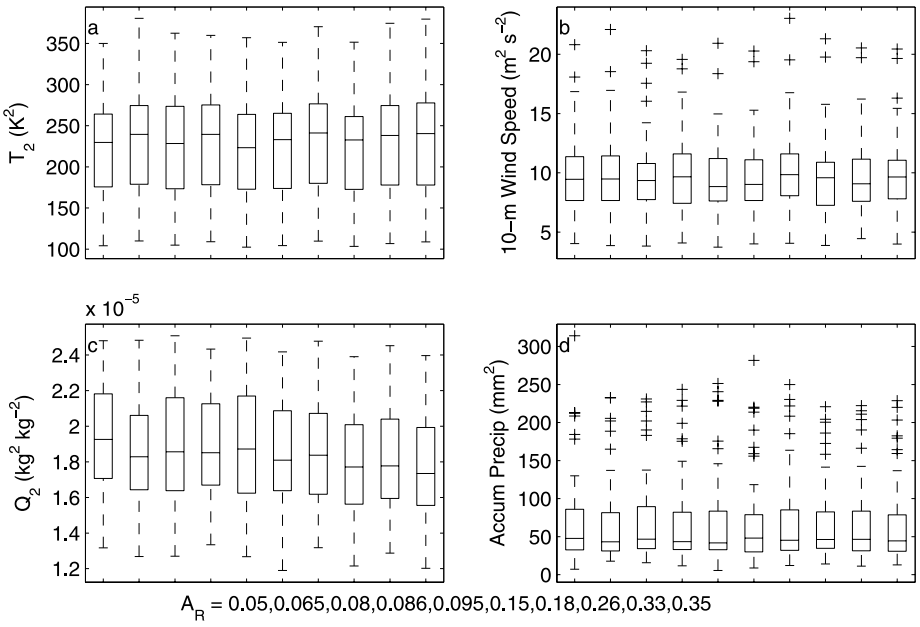

Fig. 10. Same as Fig. 8 but for spatial-variance of surface predictions.

summarized (e.g. with RMSE). The result is both an indication of distinguishability and also determination of whether a particular parameter set leads to an inferior model. In the last section, Member 1 was shown to be an outlier. Comparison to observations can show whether its error is greater than or less than the other members.

Member 1 is distinguishably deficient in predicting 10-m wind speed, but classifying any other members as deficient with these metrics would be difficult. Member 1 shows much greater surface wind-speed error than the other members (Fig. 11a) , but errors in 2-m temperature (Fig. 11b) lie within the cluster of ensemble-member errors. Use of approximately 1.5×10^5 observations at each lead time results in negligibly small 95% confidence intervals for the RMSE values, and all error levels in Fig. 11 can be considered meaningful. Wind errors are, at most, approximately 23% greater for Member 1 than for the best member. Temperature errors for the worst-performing member (not Member 1 for temperature) are, at most, approximately 10% greater than errors for the best-performing member. Differences among the other members are generally small for both variables. It is possible that another member may appear deficient if a more extensive verification were completed, but the full distributions such as those shown in Figs 8–10, would suggest that this is unlikely.

Although an explanation for the unique behaviour of Member 1 eludes us, we can rule out parameter perturbations as the cause. Member 1 is also an outlier in 10-m wind speeds for the ensemble that is directly downscaled from the global ensemble. The design of the global ensemble should prevent the skill of the directly downscaled Member 1 from being systematically different from the others. Member 1 also uses default land-surface characteristics, and it would seem highly unlikely that random perturbations applied to the other nine members all result in improved predictions.

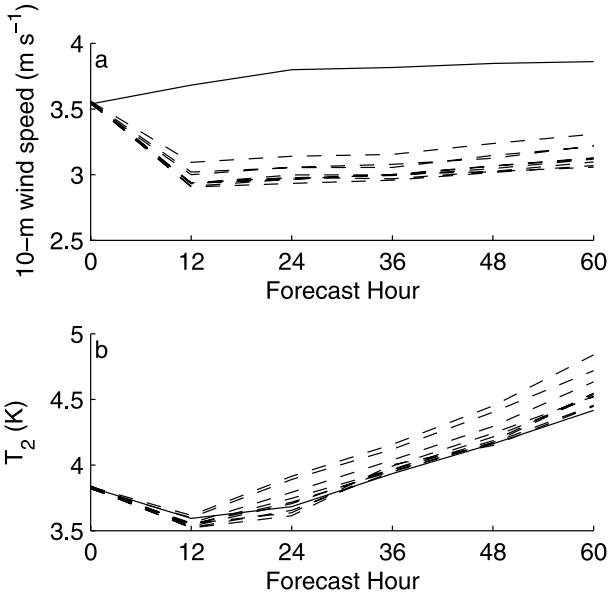

Fig. 11. RMSE of (a) 10-m wind speed and (b) 2-m temperature for each perturbed-parameter ensemble member. The solid curves show results for ensemble Member 1.

5. Local relationships

Prior sections sought systematic responses by relating distribution means (errors or state) to parameters or sets of parameters.

Those results showed that a domain-wide systematic response is lacking. Model states may exhibit stronger relationships to parameter distributions that are local in time or space, without showing a response in the mean states or mean errors. Parameter dependence can, for example, average to zero over many times and gridpoints. Here we quantify the dependence of individual model state elements on parameter values.

To evaluate strength of parameter–state relationships we choose the rank (Spearman) correlation rather than the linear (Pearson) correlation for several reason. A rank correlation between an individual state variable and individual parameter near ± 1 indicates a monotonic relationship and is not restricted to a linear relationship. Because parameter distributions are non-Gaussian and many predictions (notably wind speed and precipitation) are non-Gaussian, we do not expect a general linear relationship. The presence of additional sources of noise may hinder detection of a linear signal even if it underlies the process. Rank correlations are also resistant to outliers. One drawback to rank correlations is that they cannot be used to evaluate sensitivities as defined by Torn and Hakim (2008). High rank correlations do not guarantee high sensitivities but rather are an indication of a signal that can be exploited using transformations or non-linear methods.

To be confident in the results we test against the statistical null hypothesis that correlations are not present. For a single hypothesis test the null hypothesis can be rejected with confidence $\alpha = 1 - p$ provided by the p-value, which is the probability that a correlation is at least as high as computed. Rank-correlation p-values are computed using permutations of the data to empirically measure the probability that the correlation magnitude is equal to, or greater than, the measured correlation.

Many simultaneous hypothesis tests, such as for the single forecast of a discrete field at $n = N_x \times N_y = 11956$ gridded points here, results in a distribution of n p-values. Regions may appear to demonstrate meaningful parameter–state correlations when in truth they do not (e.g. Livezey and Chen, 1983). We therefore need a test on the correlation field, to complement tests on individual correlations.

False discovery rate (FDR) techniques attempt to test the significance of the field as a whole. The expected value of the binomial distribution of correlations (a correlation is either significant or not) is np, so here we can expect to find $p \leq 0.20$ (an easy test) in approximately 2400 of our tests if the null hypothesis of no correlation is always true. We follow the approach of Benjamini and Hochberg (1995), and the goal is to control the number of low p-values that arise by chance and suggest high confidence in the results. We will allow false detection of correlations at nominal rate q. The Benjamini and Hochberg (1995) method assumes spatial independence, but Ventura et al. (2004) and Wilks (2006) found that the algorithm performs well for spatially correlated data, and Wilks (2006) showed that q is equivalent to a field significance. In the case of spatially correlated data, it leads to a higher false detection rate than expected from the chosen value of q, and in our case identifies more instances of significant parameter–state correlations than might actually be present.

Table 3. Maximum rate of significant correlations (per cent of domain) estimated for each parameter and state variable pair, using a false-discovery rate of 20%

	T_2	Spd_{10}	Q_2	*Accum. Precip*
R	5.09	4.12	1.46	0.53
A_R	1.36	1.40	1.01	2.89
N_0	9.64	2.26	10.88	0.46
α_{SC}	3.62	0.74	1.87	0.27

Note: Maxima are found in the set of all forecast cases and forecast lead times.

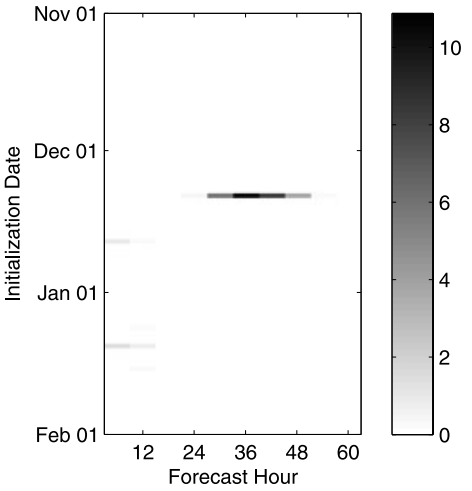

Fig. 12. Rates (per cent of domain) of significant rank correlations between 2-m water–vapour mixing ratio and parameter N_0 (microphysics scheme), as a function of forecast lead time and initialization date. 2-m water–vapour mixing ratio and parameter N_0 show the greatest rate of significant rank correlations in a forecast of any parameter–state pair in this study.

Choosing $q = 0.20$ because of the small sample size of 10 in each correlation and our desire to be permissive, we use the Benjamini and Hochberg (1995) approach for determining significance of every parameter–state correlation map. The rate of significance is then n_H/n, where n_H is the number of significant correlations in each map with n gridpoints. With $q = 0.20$, it can be interpreted as the fraction of the domain with rank correlations greater than could be expected to arise by chance less than 20% of the time. The maxima among all forecasts are shown in Table 3.

Correlations for 2-m water vapour mixing ratio and N_0 (microphysics scheme) are most meaningful under this permissive test. Figure 12 shows significance rates as a function of lead time and initialization date; the 36-h prediction initialized 0000 UTC 11 December 2008 (i.e. valid at 12 UTC 12 December)

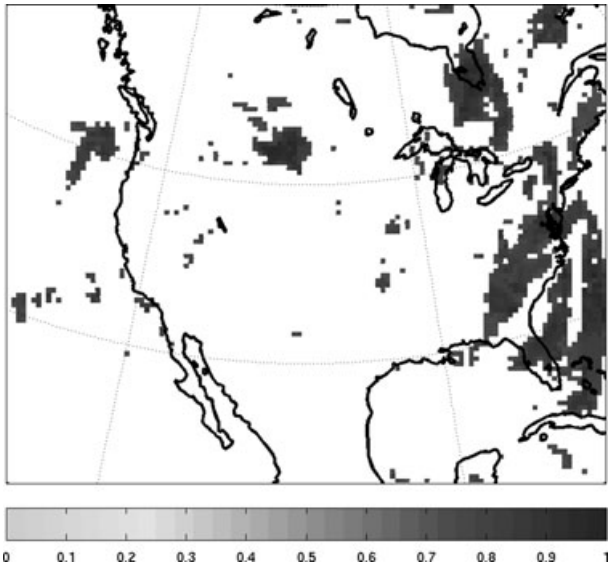

Fig. 13. Map of absolute value of 2-m water–vapour mixing ratio and parameter N_0 (microphysics scheme) rank correlation coefficients that pass the significance test, valid for a 36-h lead time at 1200 UTC 12 December 2008. This particular forecast and parameter–state pair show the greatest rates of significant correlations of any in this study.

leads to the greatest rate of significant correlations, with a rate equal to 10.88%. Figure 13 shows the absolute value of the significant correlation coefficients. White regions do not pass the significance test. The large swath of high correlation across the Eastern United States and off-shore was forecast to receive precipitation during the 24-h period between 12 UTC 11 December and 12 UTC 12 December, explaining the region of high rank correlation. Most forecasts show no significant correlations at any lead time.

Analysis with the linear (Pearson) correlation coefficient, which assumes a Gaussian distributions, justifies our choice of rank correlation coefficient. Significance rates are slightly greater for the linear correlations than the rank correlations. Correlations between 2-m water–vapour mixing ratio speed and N_0 are intermittent, but give the greatest instantaneous significance rates. Examination of significant locations and times show that the correlations are artificial, and high correlations and significance often result from the outlier Member 1.

One remaining possibility is that the ensemble of 10 members is too small to provide meaningful correlations. Analysis of 12, 100-member ensemble initialized at 00 UTC every fifth day beginning 21 November 2008 suggests that larger ensembles lead to the same broad conclusions, although slightly more optimistic. These are the same forecast cases as were used for the experiments with individual parameter perturbations, when about half of the forecast periods appear influenced by baroclinic systems. LHS is used to draw 100 parameter sets. No additional land-use perturbations or global ensemble members for LBCs were used here, so variability attributable to those sources is the same as for the 10-member ensembles.

Among those 12 forecasts, one forecast produced significance rates of 13.5% for rank correlations between R (Cu scheme) and 2-m water–vapour mixing ratio. Rank correlations between A_R (PBL scheme) and both 2-m temperature and water–vapour mixing ratio reached 10% or greater during the first 6 h of forecasts initialized 1 and 5 December 2008; cold fronts propagated through the eastern United States during both forecasts, but only the one. December 2008 forecast contained widespread precipitation. Significance rates were below 1% in all of the other forecast periods. These results suggest that a higher rate of significance may be possible with a larger ensemble. But the number of forecasts with higher rates appears small enough to be described as intermittent.

Detecting relationships between these four parameters should not require an ensemble size much greater than those used successfully in ensemble data assimilation. A large body of literature demonstrates empirically that ensembles containing tens of members can be effective, indicating that robust linear relationships between state variables can be estimated from them. Formally, the system dimensionality increases by four with the addition of four parameters that vary. Thus only a small increase in the ensemble size may be needed. A 100-member ensemble would seem large enough, especially if a suitable covariance localization could be applied.

We conclude from this analysis that intermittently significant correlations can occur within these 10- and 100-member ensembles, but most often the null hypothesis of no correlation cannot be confidently rejected. In at least a few cases, approximately 10% of the domain shows significantly correlated structures that could occur by chance less than 20% of the time. In those intermittent instances of significant correlations, Fig. 13 suggests that the parameter may need to be modelled with complicated spatial structure to be estimated.

This analysis cannot formally eliminate the possibility that significant correlations exist at a greater rate, but that they are hidden because of other noise in the ensemble predictions. A high rate of smaller-magnitude correlations, that cannot pass significance tests, exist. But qualitatively obvious relationships are difficult to find with examination of individual joint parameter–state distributions, which instead look more random and appear consistent with the field-significance test results.

These results may not extend to all parameters in a model, and to all state variables in a prediction. Here we focus on a few parameters to make analysis tractable, and on surface predictions because the Earth's surface has reasonably high-density observing networks and could support future work. Certainly other authors have found evidence of the potential for parameter estimation. But the task of finding parameters that are meaningfully correlated, and then estimating the appropriate number of degrees of freedom in each, may prove difficult.

6. Summary and discussion

This work sought to evaluate ensemble response to parameter variations, understand whether reasonable parameter variations are likely to result in an inferior model, and characterize statistical relationships between parameter values and predicted states. Unique independent sets of four parameters were drawn with a space-filling design, and 64 10-member ensemble predictions were analysed for a response to the parameter sets. Near-surface errors of the individual members were computed to determine whether any particular parameter set produced an inferior model distinguishable by large errors. We examined domain mean and spatial variances of each member to find distinguishable systematic traits of the predictions. Rank correlations between parameters and individual ensemble predictions were analysed for significance while controlling the FDR for correlation. Principal findings from the analysis can be summarized as follows.

(i) Nine of the 10 parameter sets produced models with indistinguishable distributions of forecast spatial means and variances. Member 1 was distinguishable in 10-m wind speed (Figs 8–10), but we can rule out the parameters as the cause.

(ii) Nine of the 10 parameter sets produced models with similar near-surface RMSE. Member 1 was an outlier and characterized by large 10-m wind-speed errors (Fig. 11).

(iii) Transient responses with significant rank correlations were found, suggesting that parameter estimation may be intermittently successful with linear methods, but in general parameter–state rank correlations at rates greater than could happen by chance less than 20% of the time were unusual (Figs 12–13).

(iv) Larger ensembles (100 members) may lead to slightly greater significance rates, but the broader conclusions remain intact.

We cannot rule out the existence of non-linear or noisy correlations between parameters and predictions. Weak and noisy correlations challenge current data assimilation algorithms, but do not eliminate the possibility of successful estimation. In a data assimilation system, correlations between predicted observations and model state variables will be weaker than the correlations examined here. Some challenges, such as sampling error leading to spurious correlations, are unique to ensemble filter data assimilation systems. Variational approaches may more successfully exploit weaker correlations.

Considering the possibility of non-linear relationships between parameters and forecasts could lead to clearer signal extraction from these experiments. The experiments here avoid linear relationship amongst parameters, and seek linear and monotonic relationships between parameters and predictions. Non-linear techniques proposed by Jackson et al. (2004) and Sanderson et al. (2008), for example, may extend to NWP and experiments similar to the present one. Jackson et al. (2004) propose a method based on simulated annealing to converge on parameter probability density functions for climate models, and show that convergence is possible with order 10^2–10^3 model simulations. Certainly climate*prediction*.net is capable of that many simulations. Sanderson et al. (2008) exploited climate*prediction*.net simulations to map non-linear responses, attributable to individual parameters, from thousands of simulations subject to 15 simultaneous parameter perturbations. More work is needed to evaluate the feasibility of these approaches in an NWP context.

Recently Nielsen-Gammon et al. (2010) perturbed individual parameters and found more evidence of stronger correlations than we found. They examined daytime convective PBL-scheme parameters and PBL profile predictions during August over the southern United States; PBL convection is more active in their study than in this winter-time study. Their single-parameter perturbation experiments suggest that estimating single parameter may be relatively easy. Our results suggest that simultaneous multiple-parameter estimation may be more difficult with the linear approaches used here and in Nielsen-Gammon et al. (2010). Difficulty in distinguishing responses attributable to individual parameters when multiple parameters are perturbed is consistent with Alapaty et al. (1997), who found that when multiple land-surface parameters were varied the PBL was insensitive to any one parameter. Tong and Xue (2008) and Posselt and Vukećević (2010) also found parameter estimation difficult when multiple parameters were perturbed. Results from Alapaty et al. (1997), Tong and Xue (2008), Posselt and Vukećević (2010) and Nielsen-Gammon et al. (2010) imply that single-parameter estimation experiments may be most effective; many numerical simulations are required to first find, and later individually estimate, parameters. Considering the number of parameters in a typical mesoscale model implementation, the cost may be prohibitive for many applications. Costs are even greater than they initially appear because parameter values estimated within one model, domain, and time frame may not be appropriate for another.

The lack of domain-wide systematic responses to parameters also suggests that time-dependent parameter values (e.g. that might result from imposing a stochastic process on them) are not necessary for purposes of representing parameter uncertainty in an ensemble prediction system. Although empirical evidence suggests parameter perturbations by themselves are not enough to simulate model uncertainty, they may be useful in specific implementations to augment other methods for model uncertainty representation (Bowler et al., 2008; Hacker et al., 2011). A more thorough examination of higher-order statistical moments is needed.

In a coarser-resolution global model, Rodwell and Palmer (2007) found that physics tendencies linearly varied with parameter values during the first 6 h (12 time steps in their model). The effects of a parameter perturbation, or of combining parameter perturbations, could then be predicted at short time scales. Results from the present experiments are qualitatively consistent

for all lead times greater than 6 h (we did not look at shorter times), and systematic linear responses at 6 h appear generally absent. Faster space and time scales in this mesoscale model may be the cause. For effective mesoscale parameter estimation, assimilation may need to be more frequent than 6 h.

7. Acknowledgments

This work was funded by the U.S. Air Force Weather Agency. We thank J. Rougier for suggesting Latin Hypercube Sampling as an option to draw parameter sets, and J. Dudhia and G. Thompson for suggesting potential parameters within individual physics schemes.

References

Aksoy, A., Zhang, F. and Nielsen-Gammon, J. W. 2006a. Ensemble-based simultaneous state and parameter estimation in a two-dimensional sea-breeze model. *Mon. Wea. Rev.* **134**, 2951–2970.

Aksoy, A., Zhang, F. and Nielsen-Gammon, J. W. 2006b. Ensemble-based simultaneous state and parameter estimation with MM5. *Geophys. Res. Lett.* **33**, doi:10.1029/2006GL026186.

Alapaty, K., Raman, S. and Niyogi, D. 1997. Uncertainty in the specification of surface characteristics: a study on prediction errors in the boundary layer. *Bound.-Layer Meteorol.* **82**, 473–500.

Annan, J. D., Hargreaves, J. C., Edwards, N. R. and Marsh, R. 2005a. Parameter estimation in an intermediate complexity earth system model using an ensemble Kalman filter. *Ocean Modell.* **8**, 135–154.

Annan, J. D., Lunt, D. J. and Valdes, P. J. 2005b. Parameter estimation in an atmospheric GCM using the ensemble Kalman filter. *Nonlinear Proc. Geophys.* **12**, 363–371.

Ball, F. K. 1960. Control of inversion height by surface heating. *Q. J. R. Meteorol. Soc.* **44**, 2823–2838.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57B**, 289–300.

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

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.

Cohn, S. E. 1997. An introduction to estimation theory. *J. Meteorol. Soc. Jpn.* **75**, 257–288.

Dudhia, J. 1989. Numerical study of convection observed during the Winter Monsoon Experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.* **46**, 3077–3107.

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

Fritsch, J. M. and Chappel, C. F. 1980. Numerical prediction of convectively driven mesoscale pressure systems. Part I: convective parameterization. *J. Atmos. Sci.* **37**, 1722–1733.

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., Ha, S.-Y., Snyder, C., Berner, J., Eckel, F. A., and co-authors. 2011. The U.S. Air Force Weather Agency's mesoscale ensemble: scientific description and performance results. *Tellus* **63A**, this issue.

Hong, S.-Y., Dudhia, J. and Chen, S.-H. 2004. A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Mon. Wea. Rev.* **132**, 103–120.

Hong, S.-Y. and Pan, H.-L. 1996. Nonlocal boundary layer vertical diffusion in a medium-range forecast model. *Mon. Wea. Rev.* **124**, 2322–2339.

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

Jackson, C., Sen, M. K. and Stoffa, P. L. 2004. An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model prediction. *J. Clim.* **17**, 2828–2841.

Kain, J. S. 2004. The Kain-Fritsch convective parameterization: an update. *J. Appl. Meteorol.* **43**, 170–181.

Kain, J. S. and Fritsch, J. M. 1990. A one-dimensional entraining/detraining plume model and its application in convective parameterization. *J. Atmos. Sci.* **47**, 2784–2802.

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W.-S. and co-authors. 2009. Introduction of the GSI into the NCEP Global Data Assimilation system. *Wea. Forecast.* **24**, 1691–1705.

Livezey, R. and Chen, W. 1983. Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.* **111**, 46–59.

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

Marshall, J. S. and Palmer, W. M. 1948. The distribution of raindrops with size. *J. Meteorol.* **5**, 165–166.

Moeng, C. H. and Sullivan, P. P. 1994. A comparison of shear and buoyancy-driven planetary boundary layer flows. *J. Atmos. Sci.* **51**, 999–1022.

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

Nielsen-Gammon, J. W., Hu, X.-M., Zhang, F. and Pleim, J. E. 2010. Evaluation of planetary boundary layer scheme sensitivities for the purpose of parameter estimation. *Mon. Wea. Rev.* **138**, 3400–3417.

Noh, Y., Cheon, W. G., Hong, S. Y. and Raasch, S. 2003. Improvement of the *k*-profile model for the planetary boundary layer based on large eddy simulation data. *Bound.-Layer Meteorol.* **107**, 401–427.

Posselt, D. J. and Vukećević, T. 2010. Robust characterization of model physics uncertainty for simulations of deep moist convection. *Mon. Wea. Rev.* **138**, 1513–1535.

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.

Rodwell, M. J. and Palmer, T. N. 2007. Using numerical weather prediction to assess climate models. *Q. J. R. Meteorol. Soc.* **133**, 129–146.

Sanderson, B. M., Knutti, R., Aina, T., Chirstensen, C., Faull, N. and co-authors. 2008. Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *J. Clim.* **21**, 2384–2400.

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

Sauvageot, H. and Lacaux, J.-P. 1995. The shape of averaged drop size distributions. *J. Atmos. Sci.* **52**, 1070–1083.

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.

Tong, M. and Xue, M. 2008. Simultaneous estimation of microphysical parameters and atmospheric state with simulated radar data and ensemble square-root Kalman filter. *Mon. Wea. Rev.* **136**, 1630–1648.

Torn, R. and Hakim, G. 2008. Ensemble-based sensitivity analysis. *Mon. Wea. Rev.* **136**, 663–677.

Troen, I. and Mahrt, L. 1986. A simple model of the atmospheric boundary layer: sensitivity to surface evaporation. *Bound.-Layer Meteorol.* **37**, 129–148.

Ventura, V., Paciorek, C. J. and Risbey, J. S. 2004. Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *J. Climatol.* **17**, 4343–4356.

Waldvogel, A. 1974. The N_0 jump of raindrop spectra. *J. Atmos. Sci.* **31**, 1067–1078.

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. On "field significance' and the false discovery rate. *J. Appl. Meteorol. Climatol.* **45**, 1181–1189.

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