Working Group 1: Future developments in deterministic and stochastic parametrizations

Co-chairs: Robert Pincus and Glenn Shutts

Rapporteur: Richard Forbes

Participants: Hannah Arnold, Daan Crommelin, Laurent Descamps, Jesse Dorrestijn, Henrik Feddersen, Takuya Komori, Frank Kwasniok, Hugh McNamara, Bob Plant, Petri Räisänen, David Randall, Florian Rauser, Axel Seifert, Joao Teixeira, Laure Zanna

Our group considered the desirability of including representations of uncertainty in the development of parameterizations. (By 'uncertainty' here we mean the deviation of sub-grid scale fluxes or tendencies in any given model grid box from truth.) We unanimously agreed that the ECWMF should attempt to provide a more physical basis for uncertainty estimates than the very effective but ad hoc methods being used at present. Our discussions identified several issues that will arise.

1) How can physical representations of uncertainty be developed in the context of existing techniques (SPPT, SKEB, etc.)?

Currently, the most successful technique for representing model uncertainty in ECMWF ensemble system is SPPT. This technique is the baseline, in the sense that improvements going forward will be judged relative the efficacy of SPPT.

There remains considerable scope for improvement in SPPT including calibration and a more individual, process-based variant. In particular, coarse-graining studies with IFS forecasts should make it possible to assign more credible levels of uncertainty to each of the parameterisation schemes. SPPT modifies the total parameterisation tendency and is based on a pattern generator. A natural evolution is to perturb different parametrization scheme differently, and possibly even with different patterns.

Beyond this however, the Centre should directly target the physical parameterisation schemes with respect to their inherent uncertainties. There is general agreement that deep convection, especially in the tropics, would make a good first target. There may also be opportunities to treat the uncertainty in gravity wave drag associated with unresolved orography, especially because this process is known to be very sensitive to wind profiles.

SPPT is based on pattern generators and these could also be used to modulate key parameters associated with parameterizations. Observations or process models such as CRMs can be used calibrate the spatial, temporal and structural nature of perturbations to refine the pattern generator in SPPT. It should be possible to improve on the ad hoc correlation scales associated with the three patterns in SPPT (and likewise for modulating parameters).

Recent coarse-graining studies using the IFS have been applied at the process level and have clarified the relationship between the magnitude of the uncertainty in the parameterization tendency and the mean tendency. There is growing evidence that the variance is proportional to the mean tendency. This conflicts with the underlying assumption in SPPT but is consistent with an underlying Poisson process. There also appears to be an additive component in that uncertainty exists even at zero mean tendency. Further investigation is required and EPS simulations carried out to explore the possibility that increased probabilistic skill results from using a more appropriate probability distribution function.

There was some discussion about the desirability of achieving a shallower slope in the model's energy spectrum. This could be realised through new techniques such as the vorticity confinement algorithm or other numerical methods that give non-local upscale energy transport. This may lead to improved parameterisation responses at the near grid scale.

An issue that arises when physical process uncertainty is treated individually concerns its effect within the sequential call structure of parametrization code. For instance, radiation uncertainty generally derives from cloud uncertainty and one would want to have that causality respected in the subroutine calling sequence. There was also some concern expressed about physical inconsistency in SPPT, such as the lack of surface energy flux perturbations when SPPT perturbs parametrization tendencies in the column above.

Lastly, the cultural gap between physical parametrization development and EPS development was noted as a problem, particularly for the EPS community. Traditionally, parametrization development has had deterministic NWP as its context and this can lead to problems (e.g. reduced model stability) when stochastic perturbations are generated out of the parametrization tendencies.

Recommendations

- ECMWF should test new formulations that address physical parameterisation uncertainty. For instance, the cellular automation pattern generator should be tested to perturb parameterizations.
- Tropical convection appears to be the best candidate for initial efforts, both because it has been the subject of several studies (including talks by Jakob, Majda, and Plant at this workshop) and because it impacts many other aspects of the model.
- The broader research community should be encouraged to invest effort in establishing a firm physical basis for stochastic perturbations. The most likely place to put this is directly into the physical parametrization scheme.

2) In what circumstances and for which processes do stochastic perturbations project onto the large-scale flow?

Both existing stochastic methods presently used at ECMWF to represent model uncertainty impose one or more large-scale patterns to the perturbations. This is because many algorithms that introduce noise at the grid scale do not affect forecast evolution in a significant way. But is at least one counter example in which methods operating at the grid scale do indeed change the large-scale flow. In order to be able to invest wisely in stochastic methods for representing model uncertainty we need to delineate under what circumstances and at what spatial and temporal scales do stochastic perturbations impact the large-scale dynamics.

At this stage it is unclear to what extent spatial and temporal coherence play a role in this process. In addition, it is also not clear how these effects depend on the specifics of the parameterizations, the spatial and temporal resolution, and the large-scale numerics.

It seems impossible to assess these dependencies without a concerted research effort using different methodologies and large-scale models. This effort should take place in the context of international coordinating groups such as the Working Group on Numerical Experimentation (WGNE), in order not only to entrain the overall modelling community, but also to ensure that the results of such a study are as generic and applicable as possible.

Recommendations:

- Initiate a community study, possibly within the auspices of an international group such as WGNE, to investigate this issue by testing different yet simple (to implement) stochastic perturbation methodologies (including different spatial and temporal correlations) in the context of simplified GCM simulations.

3) Implementation issues in the development of stochastic physical parameterizations

There are three existing techniques for introducing uncertainty at the model level: running multiple models, running multiple physical parameterizations within a given model, and varying parameters within a given model. The path we endorse here – developing parameterization with explicitly random elements to represent the uncertain response to a given forcing – has been less well-explored to date. Within this context, one can introduce stochastic behaviour into schemes in several ways: by varying the inputs, by perturbing parameters and assumptions within the scheme, or by introducing randomness into the scheme's response to a given forcing (e.g. through transition probabilities or finite sample sizes). Since the goal is to provide a strong physical basis for uncertainty estimates, it seems likely that each method is most appropriate for a different category of uncertainty.

Where process knowledge is high - radiation is one example – it is most sensible to perturb the inputs. That might mean sampling or integrating over a distribution, depending on the relative scales of the relevant variability and the grid size. This is a way to reflect "external" uncertainty and/or variability when the process depends strongly on the inputs.

Where process knowledge is uncertain one can perturb parameters and/or assumptions within the scheme. This approach could be used to represent uncertainty in ice habit distributions, for example.

Uncertainty and errors also arise when assumptions used to build the parameterization break down. This is so far most obvious in the treatment of deep convection: it's clear that grid sizes are now far too small to encompass a large number of deep convective elements. This can be treated by averaging the process outcome over a finite (and presumably scale-dependent) number of samples.

Some kinds of uncertainty are more ambiguous. Convection, for example, can be very sensitive to initial conditions at both the small and the large scale. It is not clear to what extent the construction of the initial ensemble samples the large-scale variability, but it's presumably important not to count this uncertainty twice. Similarly, it's important not to double-count uncertainties by adding stochastic elements and then inflating tendencies after the fact.

Enumerating various sources of uncertainty and finding appropriate ways to represent each is expected to be a significant task.

Recommendations:

- ECMWF should invest in the development of stochastic parameterizations where the physical basis for uncertainty is made explicit. We support a substantial investment, i.e. by hiring a scientist to work on the problem full time, as we expect this area to attract increasing attention in the coming years.

- ECMWF should explicitly include uncertainty treatments (likely stochastic treatments) in the development of future parameterizations for both the ensemble prediction system and the deterministic model. The latter provides a good test of the physical plausibility of the error treatment.

4) How do we make the link with fine scale models and observations?

There is a strong community working on the development of traditional parameterization, in particular in the area of clouds and convection, which is organized in programs such as GCSS. This community employs observations and high-resolution process models (CRMs, LESs) in the evaluation and development of parameterizations. We encourage the developers of stochastic parameterizations to engage in the existing activities by, for example, participating in the existing intercomparison studies. This will both enable the confrontation of new ideas in stochastic parameterization with observations and process models and help integrating what are currently somewhat separate communities.

A more comprehensive evaluation of stochastic parameterizations will require new approaches to the analysis of both observations and process model output. Observational and modelling studies presented in this workshop allowed for the quantification of the degree of stochastic behaviour of the convective response (i.e. precipitation) in relation to the large scale forcing (i.e. moisture convergence). To facilitate such studies will require the collection and storage of full 3D output at high temporal frequency of process models as well as a more comprehensive analysis of existing and future observations.

Intercomparison studies for traditional parameterizations focus on their capability of representing the mean effects of sub-grid scale processes on the large scale. For stochastic parameterization, it is necessary to evaluate whether the variability of these effects is represented realistically and to what extend these perturbations lead to realistic variability on the larger resolved scales (e.g. realistic ensemble spread). This will require the application of innovative evaluation techniques and will likely necessitate dedicated intercomparison effort for stochastic parameterizations at both the process and full model application level.

Recommendations:

- We encourage the developers of stochastic parameterization to engage in existing (intercomparison) activities wherever possible.
- The broader community should consider designing and making available dedicated CRM and observational data sets that support the development and evaluation of stochastic based parametrization (in particular clouds and convection).

We support a dedicated stochastic parametrization intercomparison project.

5) How do we represent "structural" errors in physical parametrization

A part of the uncertainty (or model error) that is not readily addressed by current methods relates to error in regions that are not targeted by the perturbed tendency, random parameters or stochastic backscatter approaches. For instance, convection parametrization may completely fail to trigger convection in some places and therefore this uncertainty will be completely missed by SPPT and random parameters which generate perturbations from the parametrization tendencies. Another structural error that might exist in convection parametrization concerns the vertical profile of convective heating which is known to play a critical role in forcing equatorially-trapped waves.

Varying parameters such as the entrainment rate might help to achieve this uncertainty in the profile of diabatic heating but SPPT would not directly address this issue.

The fact that physical parametrization is column-based also imposes structural error. An example of this is orographic gravity wave drag parametrization that assumes wave packets remain in the same grid column whereas there have been many studies recently that show that wave activity (and associated momentum fluxes) can be carried long distances from their mountain source.

Lastly, there remains the possibility of model error that is not presently recognized or understood.

Recommendation

- Address potential ‘unknown random error’ by including some additive background forcing noise to EPS perturbed forecasts (e.g. an isotropic, global vorticity forcing function)

Report of Working Group 2: Merits and drawbacks of different methods of representing model uncertainty

Co-chairs: Judith Berner and Andreas Weigel

Rapporteur: James Murphy

Participants: Jian-Wen.Bao, Tony Eckel, Normand Gagnon, Jose Garcia-Moya Zapata, Christoph Gebhardt, Chiara Marsigli, Tim Palmer, Cecile Penland, Jonathan Rougier, Erica Thompson, Claudio Sanchez, Kevin Sieck, Nils Wedi

Note: Recommendations for the international modelling community and specifically for ECMWF are shown in italics below, accompanied by supporting comments based on the group's discussions.

General Recommendations

1. Design concepts for the systematic comparison of different schemes representing model uncertainty across a range of space and time-scales, both in full and hierarchically less complex models (including small planet).

Different weather and climate prediction centres have developed different strategies for sampling model uncertainties in their forecasts. While the use of multi-model ensembles of opportunity is established in forecasting across a range of time scales, the development of more specific uncertainty methodologies (multi-parameterisation methods, stochastic parameterisation and perturbed parameter approaches) has varied between centres, and between applications. For example, there are currently several different implementations of stochastic or perturbed parameter schemes available, and the relative utility of these two approaches may vary according to the forecast time scale. Also, more work is needed to assess the benefits of combining complementary aspects of methods for sampling structural, parameter and stochastic types of uncertainty. We therefore recommend moving towards a coordinated effort to assess different approaches more systematically, across a range of prediction time scales. If international agreement could be reached on a common systematic approach, this could reduce duplication of effort in the development of multiple models and uncertainty schemes at different centres. A first step should be to agree experimental design and assessment criteria for such a comparison. This should involve the use of evidence from observations and data assimilation to diagnose the direct impacts of model perturbations (see also recommendation (10)), as well as the impacts on skill and spread in forecasts (recommendation (3)), so that the reasons for the success or failure of different schemes can be elucidated. The experience of the statistical community working on inference of complex systems from computer experiments should be used to help design appropriate experiments (e.g. Santer et al., 2003). We also recommend greater collaboration between developers of model components, to ensure that the use of alternative schemes in multi-parameterisation or multi-dynamical core experiments reflect alternatives which are judged equally credible a priori (see also recommendation (5) below), based on current understanding of the relevant physical processes.

2. The principles the different model uncertainty schemes are based upon should be stated (bottom-up).

Estimates of model uncertainty, regardless of the methodology used to obtain them, are necessarily conditional on the underlying assumptions and principles. For example, what principle prevents us from perturbing the gravitational constant g in perturbed parameter experiments? And to which degree can this principle be generalised to locate all of the coefficients in a NWP model somewhere on a spectrum that runs from g to, say, the entrainment coefficient?

The probabilistic interpretation of a multi-model ensemble requires that assumptions are made concerning the statistical properties of the individual models contributing to the ensemble. For example, are the individual models sampled from a distribution around truth, or are the individual ensemble members assumed to be "exchangeable" with the other members and the real system? Similarly, probabilistic interpretations of multi-parameter approaches are conditioned on the likelihood assigned to the choice of parameters. In addition, the outcome of stochastic parameterizations and perturbed parameter approaches depends both on the first principles within the physical process parameterizations and on the assumptions inherent to the implemented perturbation schemes. Therefore, we recommend that the basic principles applied to a model uncertainty scheme should always be explicitly stated, and that the sensitivity of the projection outcomes to these principles should be assessed in a more systematic way (see recommendation (1) above).

As developers of physical parameterizations move more towards explicitly probabilistic formulations based on first principles, these formulations should be used to emphasize the underlying physical assumptions and may help to justify, e.g. the choice of a particular parameter or stochastic perturbation.

3. The effects of different schemes generating spread should be compared and validated (top-down).

In addition to designing stochastic parameterizations from first principles within the physical process parameterizations, ensemble forecasts/climate projections should be analyzed top-down to understand the sources of ensemble spread in different model-error schemes. Emphasis should not only be on the average amplitude of spread and error but also on their spatial and temporal correlations (e.g. do unpredictable situations show larger spread than predictable situations?). Postprocessing, e.g. to create ensembles with comparable spread across different experiments, could be used to discount impacts arising purely from increased spread, allowing the effectiveness of different methodologies in reducing root mean square prediction error to be isolated (see also recommendation (8)).

4. Include uncertainty resulting from the dynamical core and physics-dynamics interactions in the assessment of model uncertainty.

In addition to uncertainty arising from the need to represent and parameterize physical processes, uncertainty arises from the truncation error of the different dynamical cores and, more importantly, interactions between the physics and the dynamics. While the difference in precision and accuracy between different dynamical cores might be small compared to typical physical parameterization errors, there is increasing evidence that the same physics parameterization might behave differently when coupled to different dynamical cores (e.g. Reed and Jablonowski, 2011). The study of uncertainty related to using different dynamical cores coupled to physics-packages is an emerging field in the "dynamical core community" and their findings should be in the awareness of the "uncertainty community", e.g. as part of the systematic intercomparison proposed in recommendation (1). A separate source of dynamical model error is associated with truncation error per se and can lead to different kinetic energy spectra in the model and potentially different predictability behavior (limited vs unlimited).

5. Models participating in a multi-model ensemble should satisfy similar standards in terms of model quality (can on short time-scales be identified by evaluation of reforecasts, e.g. BMA). Methods to identify structural similarities between models should be pursued.

Multi-model ensembles are often interpreted as sets of equally likely realizations of future weather/climate. Such an interpretation requires, amongst other considerations, that (1) the individual models are considered equally credible, and that (2) the individual models are structurally independent from each other. To satisfy (1), one must ensure that the models participating in a multi-model ensemble satisfy similar standards in terms of model quality. For predictions on short time-scales, this can be tested by the evaluation of a sufficiently large set of representative reforecasts, if available. If the models differ in their quality, such reforecasts can also be used for the computation of probabilistically meaningful model weights, e.g. by techniques such as Bayesian Model Averaging (BMA, Raftery et al. 2005). For projections on longer time-scales (e.g. multi-decadal climate projections), or when no reforecasts are available, it is less straightforward to decide whether the participating models satisfy similar standards, or how model weights should be derived. In fact, no general all-purpose metric has so far been found that unambiguously identifies how a model ranks in comparison to other models, or when a set of models should be called "similar" in terms of their quality (IPCC, 2010). Consequently, there is a need to formulate minimum standards a model is required to fulfil in order to be included into a multi-model ensemble. Probably even more problematic is the assumption of structural independence (2), which is often tacitly made but in the general case not satisfied, given that some subsets of models may share more similarities than others, for example in terms of the parameterizations and numerical schemes applied, or in terms of model components being shared, e.g. a land surface model (e.g., Masson and Knutti 2011). To enhance the reliability of probabilistically interpreted multi-model ensembles, we therefore recommend that techniques and methods are pursued that allow identifying and quantifying structural uncertainties between different models.

Specific Aspects

6. The comparison of different strategies for estimating model uncertainty should also take into account practical aspects, such as operational costs.

In principle, a systematic comparison of model uncertainty schemes (see recommendation (1) above) could involve many thousands of alternative model variants, constructed by sampling large parameter spaces or stochastic physics options in a single model, or by constructing many different structural combinations of dynamical cores and physical parameterisation schemes. In practice, operational weather forecasting centres can only entertain ensembles of limited size (given their need for timely production of high resolution forecasts), and the same is true of climate prediction centres, given their needs to use sufficient resolution to achieve credible simulations, while also including a range of earth system components. We recommend that comparisons focus on sampling model uncertainties efficiently within realistic resource constraints, with future rather than current operational set-ups in mind. Relevant approaches could involve use of short experiments (e.g. climate models run in NWP mode) and observational constraints to rule out unpromising modelling options, and deployment of appropriate statistical designs to sample plausible model variants, once identified, in a limited ensemble of forecasts.

7. Avoid confirmation biases[1] by carrying out "damn fool" experiments (e.g. vary gravitational constant).

Schemes for representing model error or improving models are typically assessed through their impacts on key emergent properties, for example the degree to which they increase forecast spread to be more consistent with forecast error, or whether a particular change to a physical parameterisation reduces a bias in a key customer-relevant variable. However, there is a risk that the experimenter's first successful attempt at achieving the desired outcome will be accepted somewhat uncritically, without having established that the outcome is being achieved for good physical reasons. This should occasionally be tested by carrying out experiments where physics settings known to be incorrect are tried in the forecast system, to test whether apparent improvements in skill, or related uncertainty estimates, could arise through a chance compensation of systematic biases, or an incorrect diagnosis of the true sources of forecast error.

8. Assess possibilities to avoid artificial clustering as introduced by multi-model ensemble systems.

In multi-model (including multi-physics) ensemble systems, artificial clustering of solutions due to shared model deficiencies among members can occur and have the potential to be very misleading if the clustering is interpreted as meaningful (i.e., physical/dynamical clustering). This can lead to problems such as over-confidence in some forecast solutions, suppression of important outliers, and biased or unrealistic probabilistic forecasts. While some such clustering is obviously synthetic and can be easily be spotted, much of it is very complex and flow dependent so cannot be sorted out. In other words: Multi-model or multi-physics approaches, while carrying the advantage of producing several forecast sub-clusters with [at least] partially independent sampling of structural model biases, give the tricky problem of interpreting the relative likelihood of the clusters. Stochastic ensembles, on the other hand, while producing a more homogeneous set of members which are easier for forecasters to use, carry the risk that all members might be wrong in the same way, hence leading to overconfident forecasts as well, but there is hope that this might be remedied as stochastic parameterizations improve. Given the current limitations of both multi-models and current stochastic parameterizations, a transitional strategy is therefore recommended:

a) Multi-model (multi-physics) ensemble members should be designed to be as diverse as possible (i.e., share fewest possible assumptions, techniques, schemes, etc., and sample as much structural uncertainty as possible)

b) Develop methodologies to distinguish between artificial clusters on the one hand and physical/dynamical clustering caused by flow-dependent growth of initial state errors on the other hand

[1] Tendency to favour information confirming preconceptions, regardless of whether the information is true.

c) Assess how statistical postprocessing may be used to detect and correct artificial clustering

d) Assess and compare the loss of prediction skill due to the artificial clustering in multi-model ensembles on the one hand, and due to the possible lack of sampling of structural errors in stochastic ensembles on the other hand.

Recommendations for ECMWF

9. Statistical postprocessing techniques based on thorough hindcast sets should be used as a benchmark strategy for assessing schemes quantifying model uncertainty.

Statistical postprocessing approaches (often referred to as calibration, recalibration or model output statistics) have been shown to represent a very efficient strategy to improve the reliability of probabilistic projections a posteriori. The efficiency and robustness of such statistical postprocessing schemes depends thereby largely on the number of independent reforecasts available. Experience has shown that bigger sets of hindcast years give better calibration statistics than larger ensemble sizes in short hindcast sets. Despite their success in improving forecast reliability, statistical postprocessing techniques are also subject to conceptual limitations. In particular, statistical postprocessing has been shown to be comparatively inefficient for improving the resolution[2] of probabilistic forecasts (at least if the ground truth used in the calibration scheme is of the same scale or coarser compared to the forecasts). Moreover, particularly for forecasts on longer time-scales such as seasonal forecasts, the statistical properties of the reforecasts may be different from those of the actual forecasts, e.g. due to the effects of climate change, or due to differences in the observational data-sets used for initialization. However, given that the effects of statistical postprocessing on forecast reliability are not easy-to-beat by other schemes quantifying model uncertainty, they should therefore be used as a benchmark strategy for the assessment of such alternative approaches (see recommendation (3)). This of course requires the availability of a sufficient number of thorough reforecasts and we recommend ECMWF continue its hindcast commitment at least at the current level to allow for such post-processing.

10. The uncertainty information should be relevant to the "best" forecast (e.g. differences in resolution between EPS and deterministic model).

ECMWF's medium-range ensemble forecast is run at lower resolution than their deterministic "best estimate" forecast. In order to be useful to forecasters, the probabilistic information in the EPS needs to be as traceable as possible to the deterministic forecast. In practice, this will depend on the forecast event of interest. For example, the deterministic model may capture some types of extreme events (e.g. an orographically-driven local downpour, or the development of a tropical cyclone) that are not represented in the EPS. It is therefore important that the EPS is designed to minimise the risk of missing such phenomena, in particular by maintaining resolution sufficiently close to the deterministic model configuration. Any outstanding structural differences in behaviour between the two systems should be clearly documented. Many forecasters like to work with specific forecasts (deterministic single forecasts, or a limited set of outcomes from a multi-model or clustered ensemble), that they can interpret and potentially adjust by applying their expert knowledge. Presentation of EPS output as a limited set of clustered outcomes may therefore encourage forecaster take-up, but it is important that

[2] Defined in the sense of the resolution term in the Brier score decomposition

the EPS is designed so that clusters represent approximately equiprobable real-world outcomes, rather than concentrations of forecast outcomes reflecting the prior sampling of similar model parameterisation options.

11. Continue and extend the use of analysis increments to look at impacts of key parameters (identified by community) / stochastic methods on systematic bias, and investigate links to forecast skill.

Running climate forecasts in NWP mode has been recognized as an important step towards isolating model problems and moving toward seamless weather and climate predictions (see recommendation (1)). In particular, initial tendencies or analysis increments (averaged over many initial dates) can be used to determine systematic biases between the model and analyses. This method provides an opportunity to confront the model with observed data and identify shortcomings in parameterizations (Klinker and Sardeshmukh, 1992, Rodwell and Palmer, 2007). ECMWF is ideally suited to test different parameters, physics packages and stochastic parameterization methods within this framework, and should do so. Key parameters or stochastic parameterization techniques could either be internally identified, or suggested by the wider community. Alternatively, the model error term computed by weak constraint 4DVar should be analyzed and its structure could aid in selecting superior stochastic or deterministic parameterization methods (e.g. in a system with model error representation one would hope the model error term is purely random and uncorrelated).

References

IPCC, 2010: *Meeting Report of the Intergovernmental Panel of Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, and P.M. Midgley (eds.)]. IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, pp. 117.

Klinker, E. and P.D. Sardeshmukh, 1992: The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci*., **49**, 608-627.

Masson, D. and R. Knutti, 2011: Climate model genealogy, *Geophysical Research Letters*, **38**, L08703, doi:10.1029/2011GL046864

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M., 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155-1174.

Reed, K. A. and C. Jablonowski, 2011: An Analytic Vortex Initialization Technique for Idealized Tropical Cyclone Studies in AGCMs. *Mon. Wea. Rev*., **139**, 689–710

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

Santner. T. J., B.J. Williams and W.I. Notz. 2003: *The Design and Analysis of Computer Experiments*, Springer, New York.

Working Group 3: Verification and evaluation of representations of model uncertainty

Co-Chairs: Tom Hamill and Carolyn Reynolds

Rapporteur: Mark Rodwell

Participants: Gianpaolo Balsamo, Roberto Buizza, Hannah Cloke, Tim Del Sole, Chris Farmer, Leo Gantner, Christian Jakob, Daniel Klocke, Mio Matsueda, Irene Moroz, Tetsuo Nakazawa, Scott Sandgathe, Stefan Siegert, Susanne Theis, Antje Weisheimer.

Below, we outline recommendations for the data sets and observations that are likely to facilitate model uncertainty verification and evaluation. We then discuss general principles of verification and evaluation relevant to model uncertainty. Finally, we highlight several new and established techniques that may be of relevance to model uncertainty evaluation, many of which were discussed during the workshop.

1) Data sets and observations

We may need additional data sets when we are developing and evaluating stochastic parameterizations (as compared to those needed when conducting process studies and improving deterministic parameterizations). In particular, we need data sets that can resolve the multiple time and space scales of interest and that will contain sufficient samples to estimate PDFs.

The most valuable observations may be ones of the fields that are produced by the parameterization scheme one is developing or evaluating. Unfortunately, such data sets may be hard to obtain, or they may be diagnosed from products of observed variables (e.g., fluxes), with concomitant increases in the uncertainty of the verification data.

As will be discussed below, we believe that a sound verification practice is to first evaluate new stochastic parameterizations at the process level, then to examine their broader impacts in the full weather and climate prediction systems. Accordingly, we provide separate recommendations for process-level and full-system testing, followed by some suggested programs where building collaborations on data sets and observations might be helpful.

Recommendations for process-level parameterization development:

1. Promote the development of nature runs from cloud-resolving models. These can be useful for the evaluation of hard-to-observe variables and may provide data that will be useful for evaluating the efficacy of stochastic parameterizations across multiple time and space scales.

2. Re-examine existing field data (GATE, BOMEX, etc.) to assess utility. Consider GEWEX/CEOP Coordinated Energy and Water-Cycle Observations Project, river discharge data (WMO GRDC), NEXRAD 4-km, hourly precipitation (radar reflectivity and rain-gauge data).

3. Utilize satellite data and derived products more widely (e.g., for cloud organization and frequency classification).

Recommendations for full-system testing:

1. Model uncertainty parameterizations such as stochastic parameterization are likely to affect the low-frequency variability of the model, such as the Madden-Julian Oscillation. These phenomena have time scales of weeks to months. To diagnose the effects on these time scales, forecasts will need to be conducted that span several seasons to years. Operational centres may need to plan for the computational resources to permit testing over such large samples, if they are not already doing so.

2. Given that testing is proposed to span several years, reanalyses will be an important data set for model initialization and as a surrogate for truth in the evaluation and verification. Given the recommendation (see below) that the uncertainty of the observation/analysis data be considered in the verification process, multiple reanalyses may be helpful. The differences between reanalysis products provide some estimate of the uncertainty in any one reanalysis. Differences between reanalyses may be particularly large for variables important for seasonal prediction, such as SST, ice cover, and land-surface properties.

3. Use satellite data more widely. As with parameterization development, satellite data is generally under-utilized.

Recommendations for teaming with other initiatives:

1. We recommend integrating or at least coordinating stochastic development with the GEWEX Cloud System Study (GCSS). This would include modifying the GCSS verification framework to include probabilistic verification.

2. EUCLIPSE (European Union Cloud Intercomparison, Process Study and Evaluation Project) may be useful to team with, as this group is collecting data from various campaigns.

3. Grey Zone project. Their cloud-resolving nature runs and focus on the development of parameterizations below 10 km is of mutual interest.

2) Principles of verification and evaluation

Verification generally refers to the process of learning about the characteristics of the forecast model by comparing forecast data with observations and/or analyses. A typical verification question might be “do forecasts from model version B resemble the observations more closely than model version A?” However, verification should also answer questions about how models go wrong as much as how they go right. Below, we provide some general guidance for the verification process that may improve the development of model uncertainty parameterizations.

Recommendations

1. Eyeball the data first. Having objective verification is desirable, but coding up a verification scheme can be time-intensive, and your own eyes may tell you what aspect of the forecast appears to be unrealistic and should be subject to the more rigorous objective testing.

2. Compare against tough reference standards. For example, a suitable test for a stochastic convective parameterization in an ensemble system may be to compare its forecasts against forecasts using another existing standard for stochastic convective parameterization, rather than against forecasts with a deterministic parameterization.

3. As verification is applied to, e.g., stochastic convective parameterizations, test them over a range of grid spacings in order to demonstrate that they are generally applicable across the range of resolutions at which the forecast model can be run in the foreseeable future.

4. Verification commonly is used to compare an instantaneous model state to observations or analyses valid at that time; this is a limited use of the verification process. In stochastic parameterization we are interested in getting the correct variability in time as well as space. Existing verification techniques can and should be applied to the time dimension as well, e.g., is the temporal change in forecast precipitation from one hour to the next rate consistent with the observed?

5. Consider verifying forecast grids averaged over different periods of time and space. A bias may be harder or easier to detect when forecasts are verified from their instantaneous state to their daily average, their monthly average, or when they are averaged over grid points or latitude bands or, say, different land-surface classifications.

6. Stochastic parameterizations, we hope, may improve the representation of larger-scale processes. Existing verification techniques can and should be applied to the verification of these physical phenomena (blocking, diurnal cycles, equatorial waves) to see if their representation has changed.

7. Incorporate observation/analysis uncertainty. Verification involves the comparison of forecast data with some surrogate for truth, typically observations or analyses. These contain errors, and the estimated error should be quantified prior to verification. Once quantified, for most probabilistic metrics there are established methods for incorporating the uncertainty of these surrogates for truth into the verification process. Similarly, it can make a difference where the “truth” comes from - own (ensemble) analysis, another centre’s (ensemble) analysis, or observations. When comparing the performance of different systems, it is important to check whether the same verification results are obtained when analyses are inter-changed.

8. Apply “falsification” concepts. When comparing one stochastic physics scheme with another, a hierarchy of falsifying tests can be conducted. At the first level, one can ask whether the stochastic physics scheme invalidates physical constraints such as producing unphysical super-saturation, unphysically large tendencies, or unphysical estimates for parameters. A top level might involve comparing proper verification scores (of process outputs, or the entire forecast system) against those of the same system with a “benchmark” stochastic physics scheme.

9. Share verification code and stochastic parameterization code using agreed-upon coding and interface standards. This will allow model developers to validate using better reference standards (see point (2) above) and may make it easier to compare results across different organizations.

10. Similarly, use standard observational databases. For example, it would be desirable for the community to agree upon and use a common cloud-resolving model and observational database to compare the various stochastic parameterizations.

11. Determine what new forecast elements should be saved to facilitate the verification of stochastic parameterizations, and save them. This includes saving sufficient model output to translate into observation space (e.g., precipitation/runoff catchment, satellite radiances).

3) Useful new techniques

Several new verification techniques have been proposed in the literature recently and/or were discussed at the workshop. Model uncertainty developers are encouraged to explore the relevance of these new techniques.

a) *Data assimilation diagnostics*. T. Del Sole discussed feasibility of performing parameter estimation through an augmented state parameter estimation technique (e.g., DelSole). This can help determine whether the parameter can be constrained reasonably at all by the data. Diagnostics of the data assimilation output, e.g., the spatial variability or temporal "jumpiness" of parameter pdfs may be informative . Failure of this may be informative about model errors, and/or whether there are compensating model errors in situations where many parameters are being estimated simultaneously. Hence, even DelSole's techniques do not yield a reasonable parameter estimate, one may learn from the process.

b) Test using new *multi-variate* verification techniques, e.g., minimum spanning tree (& Tillman Gneiting's more recent version, http://www.springerlink.com/content/q58j4167355611g1/).

c) Apply Satterfield / Szunyogh approach, which may prove useful for diagnosing how much of forecast error lies within and outside the space spanned by ensemble. http://journals.ametsoc.org/doi/pdf/10.1175/2010MWR3439.1

d) Verify covariances, correlations. Data assimilation techniques such as the EnKF can be used to use evaluate whether a particular stochastic parameterization is improving the model of covariances in an ensemble. The EnKF uses the flow-dependent covariances in the data assimilation, so a closer fit to observations provides some evidence of a more appropriate covariance model.

General recommendation for ECMWF

The ECMWF Ensemble prediction system accounts for initial uncertainty and model error. With the use of ensemble-based data assimilation, it is getting more difficult to disentangle these aspects. As models get better (for example systematic errors are decreased), there should be less need for ad-hoc strategies to inflate initial spread. In an ideal world, one should only need to perturb the observations and the model (through perturbations to the tendencies, parameters, or parametrizations). In particular, the need for singular vectors in the initial conditions should decrease. This aspect should be continuously monitored to assess progress towards the ideal situation. Comparison with the initial spread in other centre's ensemble systems would be useful. (E)

Report of WG4: Representation of model uncertainty information in data assimilation

Co-Chairs: Peter Houtekamer, Jeff Whitaker

Rapporteur: Martin Leutbecher

Participants: Harald Anlauf, Alberto Carrassi, Yannick Trémolet, Mike Fisher, Andy Majda, Hendrik Reich, Juan Ruiz, David Smith, Istvan Szunyogh, Jean-Noël Thépaut, Yonghong Yin, Shaoqing Zhang, Massimo Bonavita

In data assimilation (DA), the statistics of all sources of error need to be estimated, including the uncertainty in the model used to evolve error statistics in time. An accurate specification of model uncertainty is needed both to provide the best possible single analysis and to provide estimates of analysis uncertainty for use in initializing ensemble prediction systems. For these reasons, representation of model uncertainty is a key component of any data assimilation system.

DA systems can play a key role in developing and testing new methods for representing model uncertainty. They provide a mechanism for bringing observations to bear on the problem, and can act as a testbed for model error representation schemes. If simple methods like variance inflation are used as a baseline in ensemble DA systems, a model error scheme should improve multivariate covariances, not just spread/error consistency. The improved covariances will allow more information to be extracted from the observations and thus improve analysis quality.

When developing and evaluating model error schemes in DA systems, special care must be taken to separate other sources of error (such a mis-specification of observation error covariances) from model uncertainty.

The working group recommends that the international community focus on the following issues:

How to separate model uncertainty from other sources of error in data assimilation.

Ensemble DA systems can be tuned to provide an appropriate overall level of spread. However, it is desirable to treat different sources of unrepresented covariance separately. For example, we would expect that different methods are appropriate for treating finite ensemble size effects and errors due to deficiencies in the forecast model. Therefore, we recommend an approach that utilizes a hierarchy of idealized experimental environments, where the different sources of DA error can be controlled. Source of error in DA may include errors associated with finite sample size, errors in forward operators, errors associated with mis-specification of observation error statistics, and errors in the forecast model.

Experiments could be performed in which the “nature run” is generated by running a full NWP model that is different from the model used in the data assimilation. In addition, experiments in which the model used in the nature run and in the assimilation are the same can be used to develop methods to correct for sampling error, as well as verify that the methods used to estimate model uncertainty are not incorrectly attributing other sources of error to the model. Observing system experiments (OSE) can be utilized to check that methods for estimating model uncertainty are not sensitive to the observing network.

Developing new tools for validating schemes for representing model uncertainty in DA

For example, methods could be developed for assessing the consistency between actual and predicted error covariances in observation space.

Develop methods for estimating systematic errors in the prior and observations, so that the random component can be isolated.

Current operational systems estimate observation bias, but not forecast model bias. Methods for estimating the random component of model uncertainty often assume that the systematic component is zero.

The working group makes the following specific recommendations for ECMWF:

- Separate time scales in model error estimation for weak constraint 4DVar, so that **Q** represents the "random" component that is not correlated with the previous estimate of the model error.

  This is needed in order that the methods used in ensemble systems to estimate model uncertainty can be used also to provide **Q** for 4DVar.

- Perform a posteriori diagnostics of model error estimates produced by weak constraint 4DVar.

  This is needed in order to evaluate the physical realism of the model error estimates in collaboration with the model development group.

- Develop a fully interactive ensemble DA system (like the EnKF) in which the estimated covariances are used in the ensemble DA (and not just in a separate control analysis).

  A fully interactive system will provide a more consistent framework for developing and testing model error representation schemes.