Summary and recommendations from Working Group 3: Surface-related uncertainties and comprehensive evaluation of Model Uncertainty representations

Co-chairs: *Chiara Marsigli (DWD) and Nigel Roberts (UKMO)*

Rapporteur: *Marcin Chrust (ECMWF)*

Participants: *Ulf Andrae (SMHI), Gianpaolo Balsamo (ECMWF), Sascha Bellaire (MeteoSwiss), Clara Draper (NOAA), Sarah Keeley (ECMWF), Doug Kelley (UK CEH), Martin Leutbecher (ECMWF), Olaf Stiller (DWD), Andrea Storto (CNR)*

The aim of Working Group 3 (WG3) was to identify the key issues and recommend priorities for future research directions for ECMWF and the wider research community in two areas:

1. The methods for evaluating the representation of model uncertainty and how they could be improved, focussing on the status quo and your vision for evaluating model uncertainty representations

2. The representations of uncertainty at the surface including land, ocean, sea ice and how they could be improved

Reflecting on the status quo and your vision for the future for evaluating model uncertainty representations

Discussion points

In the context of better representing model uncertainty, we are mostly concerned about the dispersion (spread) of an ensemble and how the application of some form of stochastic representation of model error changes the dispersion. Therefore, methods that measure the dispersion, such as the variance, are very useful. But we also want to know whether the ensemble mean bias is affected, and how any changes in dispersion are related to changes in the error of the ensemble mean.

A very commonly score used for assessing ensemble performance for continuous variables (e.g., temperature, geopotential height) is the Continuous Rank Probability Score (CRPS) or fair CRPS. It is worth noting that the CRPS is sensitive to bias which means that the bias signal can creep into scorecards and potentially give a misleading view when it is primarily the variance that is of interest when attempting to account for model uncertainty. The CRPS can be decomposed into changes in mean error, changes in error variance and changes in ensemble variance (given a homogeneous Gaussian approximation) [1] and this is a useful tool for determining what has influenced the change in the CRPS when comparing two ensembles. This is different to the more conventional decomposition into reliability and resolution [2].

The issue of changes in bias is important. Some stochastic schemes change the model bias, and the current aim in ensemble forecasting is to preserve the deterministic model climate in the ensemble. Therefore, measurement of the bias is essential.

Some forecast centres apply post processing to gridded forecast fields. For example, bias or spread corrections or neighbourhood processing may be applied. This provides an additional difficulty when evaluating changes to the physics in an ensemble because the post processing and stochastic perturbations may be having similar or opposing effects, in which case the role of the post processing and combined effect also needs to be evaluated.

It is common to show graphs of ensemble skill (error of the ensemble mean) and variance or the ratio between the two. This is usually done in a domain-wide sense. However, the spread varies with geographical location and in time during the course of a forecast. Greater forecast errors should typically be found where there is greater forecast spread. It is possible for an ensemble to have a good domain-wide skill-spread relationship and not match the skill-spread very well locally. Binning the spread-skill is a useful approach for evaluating whether the ensemble variance is a reliable predictor of the mean squared error of the ensemble mean locally. For a given ensemble variance, the magnitude of the error will by quite variable when looking at individual cases and that is fine.

Observation error is still not commonly incorporated into ensemble verification. When comparing against observations, observation error includes both the error of the observation and the representativeness error (a point value is not representative of a grid cell). When comparing against analyses there is also an error in the analyses to take into account, which may need to be found using the ensemble of data assimilations (EDA). It is possible that committees are being shown score cards without knowing whether observation error is being taken into account. There can be a positive signal if observation uncertainty is accounted for, and negative in the opposite case, since the inclusion of observation error tends to increase the spread. The use of collocated observations is an approach that could help to account for observation uncertainty.

Probabilistic verification scores tend to be univariate. Multivariate probabilistic scores could be used and may be important for aspects such as interactions (e.g., land-atmosphere).

Power Spectra (including wavelets) can be used to provide a more fundamental understanding of model errors. It is expected that power laws will be broken at scales approaching the grid scale of the model and effective methods to account for unresolved processes should give improved spectra (but not right at the truncation scale). Spectra are very powerful because they can be used to both examine the properties of individual ensembles and the scales impacted by model physics perturbations.

Spatial methods (in addition to spectra) are increasingly being used to evaluate ensemble forecasts. This is particularly true for kilometre-scale ensembles that are used to forecast discontinuous variables such as precipitation or lightning or fog. Distance measures, for example the dispersion Fractions Skill Score (dFSS) and error Fractions Skill Score (eFSS) [3], can be used to find a spatial skill-spread relationship analogous to the conventional point-based skill-spread. Another benefit of the spatial approach is that it allows the upscaling of errors and dispersion to be examined. Systematically looking

at spatial dispersion in a medium range forecast is not a well-established practice, but that may change with an increase in resolution towards kilometre-scale, depending on the meteorological parameters to be evaluated. Currently there is some spatial evaluation done on the position and intensity of tropical cyclones.

Evaluation is often done for a whole domain or region when there could be additional benefit in examining the impact of stochastic physics according to geographical location, meteorological phenomena or meteorological regime. This is needed to better diagnose where and how a physics scheme may be acting. To do this may require the creation of algorithms, or use of machine learning, to define meteorological phenomena or regimes. It also requires larger sample sizes.

There are two ways of evaluating ensembles: case studies or global statistics. Both are needed. A case study provides greater insight into the influence of perturbations in a particular context, but findings cannot be generalised because it will be impossible to do enough case studies to build robust statistics. An ensemble forecast cannot be verified from a single case (or handful of cases), so determining the effect of physics perturbations on spread is fine, but an interpretation of a change in skill is much less justifiable. Analysis of case studies reflects the practice of a forecaster, which is useful because the aim of ensemble forecasts is that they are used by forecasters. When different schemes are used together, it makes it hard to evaluate their respective impact. Case studies can help because the computational expense of running several experiments is reduced. Care must be taken not to condition case studies on what was observed and hence skew results towards particular flow situations. It is better to either use a random sample of case studies (say with equal temporal sampling frequency), or to focus on particular phenomena, on the understanding the findings cannot be generalised.

There are limited resources among the community for ensemble evaluation. The sharing of evaluation methods and diagnostic tools among the community could help alleviate this problem.

Recommendations

Building on current evaluation methods

1. WG3 recommends the continued use of the CRPS (and especially the fair CRPS) but those using the CRPS should be aware that it is affected by biases and therefore the bias should also be considered independently or methods that allow the CRPS to be split into bias and spread components should be considered.

2. WG3 recommends that verification measures, especially skill-spread evaluation, should include observation uncertainty/error. This means accounting for uncertainty in both observations and analyses and including the part that is due to the representativeness of observations (i.e., comparing point observations with model grid cells).

3. WG3 encourages investigation of the use of multivariate probabilistic scores. In a dynamical system the interdependency of variables matters when thinking about the influence of stochastic perturbations on ensemble outcomes. The use of multivariate probabilistic scores

may be particularly helpful when considering interactions (e.g., between land and atmosphere).

4. WG3 encourages a greater examination of EDA methods in ensemble verification. In particular, verification of correlations or to find model uncertainty sources or exploitation of EDA methods for studying model uncertainty.

More local methods

1. WG3 recommends that the skill-spread relationship is examined more locally as well as using a domain-wide average, and that binning skill-spread is a way to do that.

2. WG3 recommends that scale-aware or spatial distance measures of spread or skill should be used more widely, especially with the move to finer resolution and an examination of less continuous variables such as precipitation or fog. This includes more use of power spectra which give an insight into the effect of the truncation scale on forecast realism.

3. WG3 encourages a greater use of partitioning in ensemble evaluation. This may be according to geographical location to better understand the effects of topography, or according to regime to better understand interactions across scales, or by meteorological phenomena to understand the effects on specific physical processes.

4. WG3 encourages more evaluation of the ensemble skill-spread for high-impact observable weather (e.g. lightning) and this should be done in a spatial sense. A feedback loop between severe weather verification and process development is desirable to better understand the origins of severe weather. It is also recognised that a focus on high-impact weather alone would not provide a dynamical picture in the same way as variables such geopotential height or temperature and should therefore be an additional exercise.

Observations and synergies

1. WG3 recognises the increasing availability of remotely sensed information, especially precipitation data from radar and satellite, and stresses the importance of making use of these data.

2. WG3 recognises the increasing availability of “non-conventional” observations and crowd-sourced observations and highlights the importance of exploiting this additional information, whilst also recognising the importance of quality control, noting the potential usefulness of machine learning for that purpose.

3. WG3 recognises that both a case study approach and a statistical approach are needed for a more complete ensemble evaluation. There is also a recognition that care is needed when using case studies for evaluating ensemble forecasts. There is benefit in engaging with

forecasters' subjective assessment and to find ways to do this in the most objective way possible.

4. WG3 would like to see more sharing of diagnostic tools among the community to help manage limited resources.

[1] Leutbecher, M, Haiden, T. Understanding changes of the continuous ranked probability score using a homogeneous Gaussian approximation. *Q J R Meteorol. Soc*. 2021; 147: 425– 442. https://doi.org/10.1002/qj.3926

[2] Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather and Forecasting*, 15(5), 559-570.

[3] Dey, S. R. A., Leoncini, G., Roberts, N. M., Plant, R. S., & Migliorini, S. (2014). A Spatial View of Ensemble Spread in Convection Permitting Ensembles, *Monthly Weather Review*, *142*(11), 4091-4107.

The representations of uncertainty at the surface including land, ocean, and sea-ice and how they could be improved

Discussion points

There was consensus that there can be a tendency to overlook the surface components and focus much more on the atmosphere.

It was noted that the dynamics of land, also sea-ice, models are very different to atmospheric models. Land models are not chaotic and are not subject to error growth in the same way as atmosphere models. Analytical solutions of the equations do exist. The difficulty is that many of the invoked parameters are uncertain or unknown (and in many cases, unknowable). The land (and sea-ice) surface is also highly heterogenous, and many of the simulated processes are non-linear. Simulating these processes using a discrete spatial unit (i.e., a model grid) will then inevitably introduce model errors, including biases in the model means.

It is important to be able to deal with different timescales. Each process has its own inherent time scales. The land surface has predictability out to the seasonal timescale, but there is also impact at short range (e.g. snow cover). It is similar for the ocean and sea ice.

There were discussions about what to perturb and why. For example, roughness length is inherently stochastic in that it is more variable than the grid box even for very fine resolution models and has a big impact on atmospheric variables. Stochasticity in snow models used within the land and sea ice models may be useful to explore and especially the parameters that control heat conduction. The recommended approach for accounting for land model uncertainty was SPP; for sea ice and ocean models this is an area of ongoing work and an intercomparison of various stochastic schemes in the ocean would be useful.

In the short-term, perturbating climatological fields like leaf area index or vegetation fraction could be low hanging fruit. These variables act on the partition of sensible and latent heat fluxes, and so perturbing them will generate reasonable ensemble spread within minimal effort. Longer term, it was recommended to perturb some selection of the parameters used in the land model parameterizations. Many of these parameters are un-observable and not well constrained. Additionally, they are determined from look-up tables based on the local land and vegetation characteristics, and the datasets for these characteristics are themselves uncertain. For example, land-use datasets are already statistically post-processed and come with prescribed uncertainty. It was stated that it is known the land-use is often misrepresented at pixel level, therefore we may design suitable uncertainty distributions for the parameters linked to land-use. Additionally, we should request uncertainty quantification of the parameter estimates that are provided for use in the surface model.

It would be beneficial for scientists working on perturbations in the Planetary Boundary Layer (PBL) and those working on surface perturbations to work more closely.

For the land and sea ice models the sub grid variability matters, and the heterogeneity should be represented stochastically. A better idea of variability can be gained by coarse graining. The use of high resolution satellite skin temperature could help to determine what is the real structure and variability at fine scales. Potentially non-conventional/crowd sourced observation can be useful in improving understanding of land surface uncertainties and variability. It is an advantage to use surface tiles in a model, but the fractions are uncertain.

The role of data assimilation was discussed. It was noted that first guess departures can be used to adapt the model parameters in order to reduce model biases. Perturbing parameters is an obvious thing to do in an EDA context and parameter estimation could be included in the data assimilation process. If only slow processes are perturbed there is unlikely to be enough spread on the assimilation window time scale. Fast processes could follow a similar approach as for the atmosphere, e.g., SPP; perturbing fast processes has less chance of affecting the bias. From the ocean side, perturbing observations provides a larger spread than what is introduced by stochastic physics; for seasonal prediction the combination of the two should be ideal.

The move to kilometre-scale models is inevitable and that brings changes to the way the land surface is represented. Urban representation becomes essential, the dominant vegetation becomes more representative of a grid cell with less sub-grid heterogeneity of vegetation. It becomes important to distinguish between low- and high-rate vegetation. Synergies having viable and physically consistent land surface perturbations could be particularly important for the PBL.

Given the non-linear nature of the models, applying a scheme to generate spread in a land model ensemble is expected to change the ensemble mean land states (and fluxes), so that the control (un-perturbed) model calibration is no longer optimal. It is unclear how to resolve this. On the one hand, if the scheme used to generate the ensemble spread adequately represents the true model errors, the change in ensemble mean (compared to an un-perturbed control run) is a measure of the model bias associated with the non-linear model response to its uncertain parameter set, and this

information should not be discounted. On the other hand, it is not practical to expect model developers to calibrate land models in ensemble mode, and then maintain separate un-perturbed and ensemble calibration parameter sets. Current practice is to limit the magnitude of the ensemble spread near land to minimize any changes in the ensemble mean. However, this results in a substantially under-dispersed ensemble.

Recommendations

WG3 recommends that more work is undertaken in stochastic perturbations of earth-system components interacting with the atmosphere at the surface and can see several areas in which this could be beneficial for ensemble performance. Work on representing model uncertainty at the surface is encouraged.

What to perturb

1. WG3 particularly encourages work on stochastic perturbations of:
    a. Roughness length, which can have a big impact on atmospheric models
    b. Cryosphere – perturbation of parameters that control snow density is a good way of introducing stochasticity to snow models
    c. Climatological fields such as leaf area index
    d. Un-observable, or otherwise highly uncertain, model parameters.
    e. Sea surface temperature (SST)
    f. Sea ice fraction/thickness
    g. Wind stresses over the ocean

Note that (e) and (f) may be suitable for atmospheric models that are not coupled to ocean/sea-ice models.

WG3 also recommended setting the magnitude of the applied perturbation distributions to represent the local sub-grid variability, rather than using the same perturbation distribution globally.

Methods

1. WG3 recommends the application of coarse graining to provide guidance on sub-grid variability. Perturbations may be too large if the aim is just to correct mean errors.
2. WG3 recognizes that the main method used for surface perturbations is Stochastic Parameter Perturbations (SPP) and this is generally applied with the same perturbation distribution everywhere.
3. WG3 notes that the perturbation schemes are likely to affect the ensemble mean and hence require a different calibration. This is an important aspect that requires further thought.

Data assimilation

1. WG3 encourages the use of first guess departures within land surface data assimilation to guide how to adapt model parameters in order to reduce model biases
2. WG3 encourages perturbing parameters in an EDA context.

Synergies, enhance collaboration

1. WG3 recommends that the community compile a list of parameters that are known or can be derived from physical considerations and a list of parameters that are unknown, which would entail more freedom in tuning/perturbing.
2. WG3 identifies the Planetary Boundary Layer (PBL) as a common joint focus area for the land surface, also sea-ice, and atmospheric research and modelling communities and agrees that there would be benefit from considering perturbations of PBL parameters and surface perturbations together rather than independently.

Vision

1. WG3 recommends that future land surface models should be designed with stochasticity in mind. The current situation is that stochastic perturbations are added to fundamentally deterministic models and that has limitations.

2. WG3 recommends requesting uncertainty quantification of parameter estimates that are provided for use in the surface model (and 2D-fields).

3. WG3 recommends exploring the use of non-conventional/crowd-sourced observations as a potential useful means of improving understanding of land surface uncertainties.

4. WG3 recognises the move towards kilometre-scale models. This brings less sub-grid variability, but a greater need to represent low- and high-rate vegetation and a greater importance of representing urban areas.

5. WG3 recognises that uncertainty knowledge is under-utilised in the ocean community and recommends fostering more links.