Model Uncertainty Intercomparison Project: Discussion of Suggested Protocol

Hannah Christensen[*]

8th February 2019

1 Introduction and Motivation

Stochastic parametrisations are widely used by weather and seasonal forecasting centres to represent model uncertainty. Three approaches in particular have become widely used. In part this is due to their beneficial impacts on forecast spread, improving the reliability of forecasts, though the ease of implementation and generalisability to different models likely also plays a role. The first approach is the 'Stochastically Perturbed Parametrisation Tendencies' (SPPT) scheme, which addresses model uncertainty due to the parametrisation process (Buizza et al., 1999; Palmer et al., 2009; Yonehara and Ujiie, 2011; Bouttier et al., 2012; Sanchez et al., 2016; Berner et al., 2015; Christensen et al., 2017; Davini et al., 2017). It does this by perturbing the sum of the parametrised physics tendencies using multiplicative noise:

$$\mathbf{T}_X = \mathbf{D}_X + (1+e)\sum_{i=1}\mathbf{P}_{i,X} \tag{1}$$

where $\mathbf{T}_X$ is the total vector tendency in X, as a function of model level at a particular spatial grid point. $\mathbf{D}_X$ is the vector tendency from the dynamics, $\mathbf{P}_{i,X}$ is the vector tendency from the ith physics scheme, and e is a zero mean random perturbation. The second approach is the 'Stochastic Kinetic Energy Backscatter' (SKEB) scheme (sometimes called the Stochastically Perturbed Backscatter Scheme: SPBS) (Shutts, 2005; Berner et al., 2009, 2012; Tennant et al., 2011; Sanchez et al., 2016). The SKEB scheme is designed to represent a physical process absent from deterministic models, namely the upscale transfer of kinetic energy from small to large scales. This counteracts the kinetic energy loss at small scales from excessive dissipation in numerical integration schemes. This backscatter is achieved by randomly perturbing the streamfunction at large scales, with an amplitude modulated by the subgrid dissipation rate. The final approach uses expert elicitation to select uncertain model parameters, and provide a bound on the possible values of these parameters (Bowler et al., 2008; Ollinaho et al., 2013, 2017; Jankov et al., 2017). Under this 'Random Parameter' (RP) approach, the value of the selected parameters is stochastically varied within this range to account for uncertainty in the forecast due to the processes represented by these uncertain parameters.

Despite the widespread use of these three approaches, little work has been carried out to assess the realism of these representations of model error, and in particular, how the characteristics of random model error differ between different models. It is therefore not known whether approaches such as the three outlined above are optimal for all operational forecast models.

A promising approach to characterise and understand model error is the use of coarse-graining studies (Shutts and Palmer, 2007; Shutts and Pallares, 2014; Dorrestijn et al., 2013; Porta Mana and Zanna, 2014; Bessac et al., 2019). This approach takes high-resolution atmospheric simulations, in which the key processes of interest are resolved, as a proxy for the real atmosphere. The high-resolution

[*]Atmospheric Oceanic and Planetary Physics, University of Oxford, OX1 3PU

simulation is coarsened to the resolution of the forecast model of interest, and the difference between the low resolution forecast and the coarsened high-resolution simulation is considered the model error that a stochastic parametrisation seeks to represent. Coarse-graining studies have traditionally been computationally expensive and complicated to carry out. A modelling centre must produce both a high resolution reference simulation and a low resolution forecast. To measure the instantaneous statistics of model error, these forecasts must be reinitialised at (close to) every time step of the low resolution forecast model. Recently, an alternative approach has been proposed, whereby existing high-resolution datasets can be used to derive the initial condition and forcing files to drive a low-resolution Single Column Model (SCM) (Christensen et al., 2018; Christensen, 2019). This allows centres that are unable to produce high-resolution simulations to use a coarse-graining approach to characterise model uncertainty.

At the joint 33rd Meeting of the Working Group on Numerical Experimentation (WGNE) and the 4th meeting of the Predictability, Dynamics and Ensemble Forecasting (PDEF) Working Group in Tokyo (October 2018), a proposal was put forward for a co-ordinated activity to evaluate model error across a number of forecast models. This document outlines a potential protocol for such an activity.

2 Outline

The intercomparison project will consist of four key stages.

1. Produce limited area high resolution simulations to use as benchmarks. Validate fidelity of these simulations.

2. Coarse-grain these simulations to a chosen common resolution

3. Use the coarse-grained dataset to drive a number of SCM

4. Analyse model error characteristics through comparison of SCM with coarse-grained benchmark simulations.

There are a number of key decisions that need to be made prior to launching the project, which will depend on the priorities of participating institutes. These are outlined in the flow-chart in Figure 1. Interested parties can indicate their willingness to participate in any of the four stages. There is no requirement to produce both a high-resolution simulation and to perform low-resolution forecasts with the same model.

2.1 Stage 1: High resolution simulations

A number of options were discussed at the WGNE/PDEF meeting. It was decided that new simulations be limited area, for ease of production, and that simulations shall focus on ocean regions, to simplify the coarse-graining protocol. It is proposed that a number of different regions be analysed, depending on available resources. In order of priority, these regions are:

1. **Indo-Pacific Warm Pool**, e.g. 20°S–20°N, 42–177°E (Cascade domain). This highly studied region was the focus of the Year of Tropical Convection (YOTC) and Years of the Maritime Continent (YMC) initiatives. It includes a region covered by the Darwin Radar, for validation. The period and domain of simulation will be chosen to match an Intensive Observations Period (IOP) such as those that have already been carried out under the YMC, or the upcoming TerraMaris field campaign (November 2019 – March 2020).

2. **North Atlantic**, 30°N–66°N, 60°W-5°W. An extratropical region is selected to contrast with the tropical Indo-Pacific region. The North Atlantic is a region of key interest for studying the jet stream and storm tracks important for forecasting European weather. The period and domain

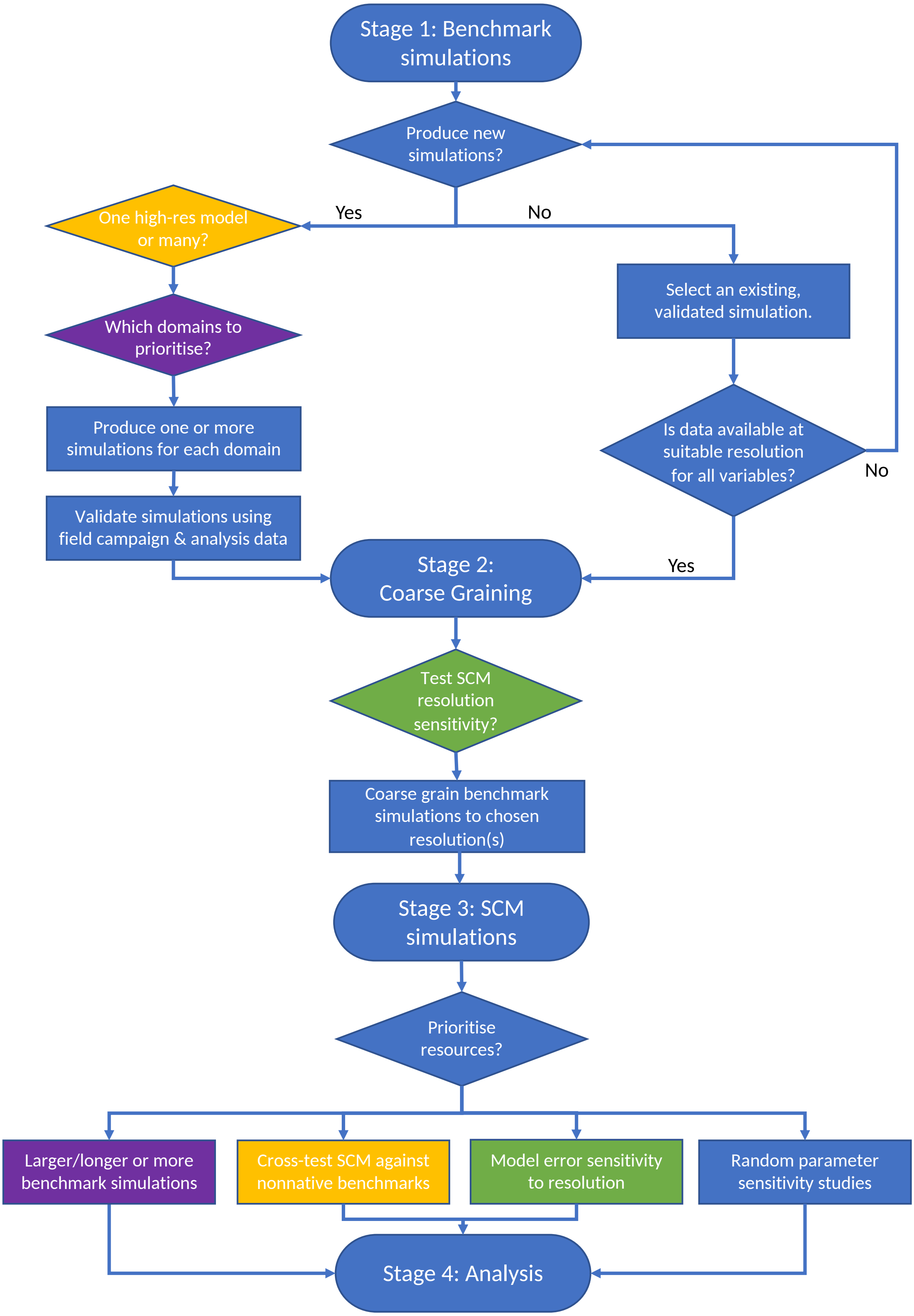

Figure 1: Flowchart showing the main stages for the intercomparison activity, and associated key decisions. Colours other than blue indicate related decisions/activities

of simulation will be chosen to coincide with the North Atlantic Waveguide and Downstream Impact Experiment (NAWDEX) field campaign, September-October 2016, to allow for model validation.

3. **Summer Arctic** e.g. North of 66°N. Recent years have seen a rapidly growing interest in polar regions, as demonstrated by the World Meteorological Organisations Polar Prediction Project, with its flagship activity: the Year of Polar Prediction (YOPP). Numerical weather prediction models have historically performed poorly in polar regions, but the increase in observational data, partly fuelled by YOPP, allows us to focus on model performance in those regions. The period and domain of simulation will be chosen to coincide with a YOPP IOP, July-September 2018, to allow for model validation.

4. **Southern Indian Ocean** e.g. 60°–30°S, 20–120°E. The Southern Indian Ocean is a large region with zonally symmetric storm tracks, providing an idealised test-case for assessing stochastic parametrisation.

5. **Tropical Atlantic** e.g. 5°N–20°N, 65–15°W. The Tropical Atlantic has been used as a testbed for understanding the role of shallow clouds in climate response, and the feedbacks between clouds and circulation. A large number of research flights have been performed over the tropical Atlantic. The period and domain of simulation will be chosen to coincide with the NARVAL2 field campaign – August 2016.

There is also the option of making use of existing high-resolution datasets, thereby skipping over this first step. For example, the ESiWACE-DYAMOND project has produced a suite of global high-resolution simulations which are publicly available. This has the benefit of not duplicating effort, and accelerating progress within this activity, though we then have to work within the framework decided by those who produced the high resolution datasets. For example, the DYAMOND protocol requires 3D fields to be stored only 3-hourly and only up to 20km. Even if an existing global simulation were used, for practical reasons the coarse-graining analysis would still be carried out separately over one or more limited area domains. Potential existing simulations include (n.b. incomplete list, and individuals have not been approached):

- Global: ESiWACE-DYAMOND simulations. Nine modelling groups have committed to produce sub-5km simulations for 40 days.
- Domain 2: 'NAWDEX-AMIP' simulations with MetUM in operational NWP configuration (Claudio Sanchez)
- Domain 3: YOPP 'Frontier Experiments' at 2.5 km resolution or higher (ECCC, Met-Norway, Meteo-France)
- Domain 5: ICON HErZ NARVAL-II HD(CP)^2 simulations (Daniel Klocke, DWD and Matthias Brueck, MPI-M) 1.2–2.5km resolution.

All new or existing simulations are expected to meet the following requirements, adapted from the DYAMOND protocol:

- Initialisation will be from a common (e.g. ECMWF or Met Office) atmospheric analysis, and run for the agreed period for each region using specified sea-surface temperatures.
- The high resolution simulations should be convection permitting (i.e. a resolution of 1–5 km) and should not include parametrised deep convection. The vertical domain should extend above the tropopause (e.g. 25-30 km or higher).

- Models are expected to represent atmospheric processes reasonably well, and will be verified before use in the coarse-graining analysis by comparison with observational data and analysis products.
- Model output will be stored hourly on model levels for the prognostic variables (U, V, T, q, q_l, q_i, cloud fraction) and hourly at the surface for boundary forcing (including surface sensible and latent heat fluxes, skin temperature, SST).
- A spin-up period will be discarded from the beginning of each simulation, and the remainder used for analysis.

A point to consider here is that if we produce our own high-resolution simulations, then we can also produce a sister simulation at a lower resolution (or perhaps, at a range of lower resolutions). This has been indicated as of particular interest to the HIWeather and Waves2Weather consortia. They have developed a diagnostic for assessing upscale error growth, and attributing grown to the representation of particular processes of interest. Other groups may also find these paired runs useful, and they would complement the SCM analysis allowing assessment of the feedback from the errors onto the dynamics.

2.2 Coarse Graining

The coarse-graining must be carried out separately for each high-resolution simulation. To combine the coarse-graining procedure with the low-resolution forecast model, we adapt the methodology described in Christensen et al. (2018). A Single Column Model (SCM) will be used to integrate forward the equations of motion in each coarse-scale grid column.

The coarse-graining methodology is detailed in Christensen et al. (2018) and Christensen (2019). An overview is reproduced here for ease of reference. The low-resolution forecast model's grid is used to define the latitude and longitude co-ordinates that make up the coarse-scale SCM grid. The fields from the high-resolution benchmark simulation are coarsened onto the low resolution grid using local area averaging. This allows for high-resolution grid boxes to contribute a fractional component to several coarse-resolution grid boxes:

$$\overline{\psi}_{n,k} = \sum_f W_{n,f}\psi_{f,k} \tag{2}$$

where ψ_f denotes the field on the fine grid and $\overline{\psi}_n$ denotes the field on the coarsened grid. The coarse (fine) grid box is identified by the index n (f). $W_{n,f}$ indicates the fraction of fine grid box f within coarse grid box n, and the vertical level of the field is indicated by index k.

Both the fine- and coarse-resolution datasets are defined on model levels, and interpolation must also be performed in the vertical. We choose to perform vertical interpolation second, after first averaging horizontally across each model level. The first field to be coarsened in this way is the surface pressure. The low-resolution surface pressure field is used to define the pressure on the SCM vertical levels, e.g. using the SCM's hybrid height coefficients. The coarse-grained benchmark dataset is then interpolated, logarithmically in pressure, from the benchmark model levels to the SCM model levels (Christensen et al., 2018). Any data required by the SCM above the benchmark model top will be taken from an operational analysis dataset. Finally, a 9-point gaussian smoother is applied to all initial condition fields after coarse graining. This removes small scale features present in the benchmark simulation that are unresolved on the low resolution grid, and which therefore appear as grid-point noise.

The advected tendencies of the prognostic variables (T, U, V, q) are calculated along SCM model levels from the coarsened fields:

$$\mathrm{adv}(\psi)|_{n,k} = -\overline{\mathbf{u}}_{n,k} \cdot \overline{\nabla}_k(\overline{\psi_{n,k}}) \tag{3}$$

for variable ψ. A centred finite difference scheme is used to estimate the vector gradient in ψ before the dot product is taken with the coarse-grained vector wind field, $\overline{\mathbf{u}}_{n,k}$. Any other required forcing, such as a geostrophic wind forcing or vertical velocity forcing, are also evaluated using the coarse-grained fields: see Christensen et al. (2018) for more details.

The constant boundary fields required by the SCM are taken from the relevant global model at the appropriate resolution, ensuring the SCM has the same boundary conditions as the global model. Interactive land surface processes are turned off in the SCM, and replaced with time varying latent and sensible heat fluxes from the benchmark simulations (provided this functionality is available in the SCM).

An open question remains as to whether to coarse-grain in time as well as space. This is a somewhat philosophical question as to what does a grid point field represent. Previous testing has only been carried out *without* coarse-graining in time, but rather using instantaneous coarse-grained fields.

NCL software has been published which coarse-grains a high-resolution MetUM simulation to produce forcing fields for the OpenIFS SCM: https://github.com/aopp-pred/cg-cascade. The software has been tested over the MetUM 'Cascade' dataset, covering the Indo-Pacific region outlined above. Coarse-grained Cascade datasets are archived at the NERC Centre for Environmental Data Analysis (CEDA): http://catalogue.ceda.ac.uk/uuid/bf4fb57ac7f9461db27dab77c8c97cf2.

2.3 Low-resolution SCM forecasts

Each different SCM will require an independent set of coarse-grained forcing files due to differences in the model set-up (e.g. different vertical grids, different prognostic variables, ...). A first step is to assess how different the input files will need to be, to understand whether the coarse-graining procedure will need to be carried out independently for each SCM (worst case scenario) or if the input files for one model can be transformed to those needed by the other models.

A SCM integration will be initialised once an hour for each grid box of the coarse-grained benchmark simulation. If the SCM exhibits a marked spin-up period over the first few timesteps (such as the IFS SCM: see Christensen et al. (2018)), then each simulation will two hours, to the nearest number of integer SCM timesteps. The first hour of each SCM simulation will be discarded, and the second hour considered for analysis. This is to focus on error statistics relevant to the bulk of the model simulation. If no spin-up period is observed then each SCM simulation will last one hour. The SCM will not be nudged to the coarse-grained benchmark fields, but will be allowed to evolve freely.

Note that given the proposed size of the high-resolution domains, the SCM will be run independently over many thousands of spatial grid boxes, and initialised hourly for the duration of the benchmark simulation. Doing this efficiently and with minimal user input is a computational challenge. However, python software has been developed to facilitate this procedure and will be available for this intercomparison activity:

- 'scmtiles': Python software to deploy many independent SCMs over a domain. https://github.com/aopp-pred/scmtiles
- 'openifs-scmtiles': Python software to deploy the OpenIFS SCM using scmtiles. https://github.com/aopp-pred/openifs-scmtiles

Note that 'openifs-scmtiles' will need to be adapted by each modelling centre to meet the needs of their own SCM.

3 Analysis and Outcomes

- Compare the characteristics of systematic 'model error' across a range of models, i.e. the difference between the forecast SCM and the coarse-grained benchmark simulation.

- Use knowledge about fidelity of the high-resolution simulations to ascribe this to deficiencies in the SCM, or to biases in the benchmark simulation.
- Attribute errors to specific model deficiencies using parametrised tendency information.
- Assess how systematic errors are affected by different geographic regions/ flow regimes.
- Depending on priorities, consider dependency of model error on resolution across a range of models.

- Compare the characteristics of random model error across a range of models.
  - Characterise degree of state dependency of random error (e.g. within SPPT framework: see Christensen (2019)).
  - Assess stochasticity in other processes of interest specified by partners. E.g. for convective processes, consider variability in CAPE, CIN or updraft velocity as diagnosed by the SCM parametrisation. Discussion ahead of time will ensure the relevant benchmark simulation and SCM outputs are archived by all centres.
  - To assess RP will require a set of SCM simulations in which parameters are perturbed for each SCM simulation. Searching over a large parameter space will be computationally expensive, but this could nevertheless be chosen as a priority.
  - Assess how random errors are affected by different geographic regions/ flow regimes.
  - Depending on priorities, also assess dependency on resolution.
  - N.b. It is not easy to assess SKEB in this framework, because there is no feedback from the SCM onto the large-scale state, and therefore no upscale cascade.

4 Next steps and open questions

- Which modelling centres are interested in participating in this activity?
- What existing high-resolution simulations are available, and are they suitable for our purpose?
- Which modelling groups have the capability and inclination to produce new high-resolution simulations for this activity? How many days of simulation would be possible (in total) and at what resolution?
- For groups interested in running a SCM, what initial condition files and forcing fields are needed by their model and on what vertical grid?
- What are the key priorities of the interested groups? How would groups choose to prioritise their time and resources from among the range of possible research paths?
- Where will the high-resolution data be stored? Can it be moved to where each SCM will be run? Or can other groups run their SCM where the data will be stored?
- Are there other research questions that we would like to consider? For example, are there other specific stochastic approaches that could be assessed (e.g. the Plant-Craig scheme)
- Any other feedback on the proposed framework?

References

J. Berner, G. J. Shutts, M. Leutbecher, and T. N. Palmer. A spectral stochastic kinetic energy backscatter scheme and its impact on flow dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.*, 66(3):603–626, 2009.

J. Berner, T. Jung, and T. N. Palmer. Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Climate*, 25(14):4946–4962, 2012.

J. Berner, K. R. Fossell, S.-Y. Ha, J. P. Hacker, and C. Snyder. Increasing the skill of probabilistic forecasts: Understanding performance improvements from model-error representations. *Mon. Weather Rev.*, 143:1295–1320, 2015.

J. Bessac, A. Monahan, H. M. Christensen, and N. Weitzel. Stochastic parameterization of subgrid-scale velocity enhancement of sea surface fluxes. *Mon. Weather Rev.*, 2019. Submitted.

F. Bouttier, B. Vié, O Nuissier, and L. Raynaud. Impact of stochastic physics in a convection-permitting ensemble. *Mon. Weather Rev.*, 140(11):3706–3721, 2012.

N. E. Bowler, A. Arribas, K. R. Mylne, K. B. Robertson, and S. E. Beare. The MOGREPS short-range ensemble prediction system. *Q. J. Roy. Meteor. Soc.*, 134(632):703–722, 2008.

R. Buizza, M. Miller, and T. N. Palmer. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. Roy. Meteor. Soc.*, 125(560):2887–2908, 1999.

H. M. Christensen. Constraining stochastic parametrisation schemes using high-resolution simulations. *Q. J. Roy. Meteor. Soc.*, 2019. submitted.

H. M. Christensen, J. Berner, D. Coleman, and T. N. Palmer. Stochastic parametrisation and the El Niño-Southern Oscillation. *J. Climate*, 30(1):17–38, 2017.

H. M. Christensen, A. Dawson, and C. E. Holloway. Forcing single column models using high-resolution model simulations. *JAMES*, 10, 2018. doi: https://doi.org/10.1029/2017MS001189.

P. Davini, J. von Hardenberg, S. Corti, H. M. Christensen, S. Juricke, A. Subramanian, P. A. G. Watson, A. Weisheimer, and T. N. Palmer. Climate sphinx: evaluating the impact of resolution and stochastic physics parameterisations in the ec-earth global climate model. *GMD*, 10(3):1383–1402, 2017.

J. Dorrestijn, D. T. Crommelin, A. P. Siebesma, and H. J. J. Jonker. Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data. *Theor. Comp. Fluid Dyn.*, 27(1–2):133–148, 2013.

I. Jankov, J. Berner, J. Beck, H. Jiang, J. B. Olson, G. Grell, T. G. Smirnova, S. G. Benjamin, and J. M. Brown. A performance comparison between multiphysics and stochastic approaches within a north american RAP ensemble. *Mon. Weather Rev.*, 145:1161–1179, 2017. doi: https://doi.org/10.1175/MWR-D-16-0160.1.

P. Ollinaho, M. Laine, A. Solonen, H. Haario, and H. Järvinen. Nwp model forecast skill optimization via closure parameter variations. *Q. J. Roy. Meteor. Soc.*, 139(675):1520–1532, 2013.

P. Ollinaho, S.-J. Lock, M. Leutbecher, P. Bechtold, A. Beljaars, A. Bozzo, R. M. Forbes, T. Haiden, R. J. Hogan, and I. Sandu. Towards process-level representation of model uncertainties: Stochastically perturbed parametrisations in the ECMWF ensemble. *Q. J. Roy. Meteor. Soc.*, 143(702): 408–422, 2017.

T. N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. J. Shutts, M. Steinheimer, and A. Weisheimer. Stochastic parametrization and model uncertainty. Tech. Mem. 598, European Centre for Medium-Range Weather Forecasts, Shinfield park, Reading, 2009. URL `http://www.ecmwf.int/en/elibrary/technical-memoranda`.

Pier Gian Luca Porta Mana and Laure Zanna. Toward a stochastic parameterization of ocean mesoscale eddies. *Ocean Modelling*, 79:1–20, 2014. doi: 10.1016/j.ocemod.2014.04.002.

C. Sanchez, K. D. Williams, and M. Collins. Improved stochastic physics schemes for global weather and climate models. *Q. J. Roy. Meteor. Soc.*, 142(694):147–159, 2016.

G. Shutts. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. Roy. Meteor. Soc.*, 131(612):3079–3102, 2005.

G. J. Shutts and A. C. Pallares. Assessing parametrization uncertainty associated with horizontal resolution in numerical weather prediction models. *Phil. Trans. R. Soc. A*, 372(2018), 2014.

G. J. Shutts and T. N. Palmer. Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization problem. *J. Climate*, 20(2):187–202, 2007.

W.J. Tennant, G. J. Shutts, A. Arribas, and S. A. Thompson. Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Weather Rev.*, 139: 1190–1206, 2011. doi: https://doi.org/10.1175/2010MWR3430.1.

H. Yonehara and M Ujiie. A stochastic physics scheme for model uncertainties in the JMA one-week ensemble prediction system. Technical Report 41, CAS/JSC WGNE Research Activities in Atmospheric and Oceanic Modelling, 2011.