Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

Methods of investigating forecast error sensitivity to ensemble size in a limited-area convection-permitting ensemble

Ross Noel Bannister[1], Stefano Migliorini[1,2], Alison Clare Rudd[1,3], and Laura Hart Baker[1]

[1]Dept. of Meteorology, Univ. of Reading, Earley Gate, RG6 6BB, UK.
[2]Now at Met Office, FitzRoy Road, Exeter, EX1 3PB, UK.
[3]Now at Centre for Ecology and Hydrology, Wallingford, OX10 8BB, UK.

Correspondence to: Ross Bannister (r.n.bannister@reading.ac.uk)

Abstract. Ensemble-based predictions are increasingly used as an aid to weather forecasting and to data assimilation, where the aim is to capture the range of possible outcomes consistent with the underlying uncertainties. Constraints on computing resources mean that ensembles have a relatively small size, which can lead to an incomplete range of possible outcomes, and to inherent sampling errors. This paper discusses how an existing ensemble can be relatively easily increased in size, it develops a range of standard and extended diagnostics to help determine whether a given ensemble is 'large enough' to be useful for forecasting and data assimilation purposes, and it applies the diagnostics to a convective-scale case study for illustration. Diagnostics include the effect of ensemble size on various aspects of rainfall forecasts, kinetic energy spectra, and (co)-variance statistics in the spatial and spectral domains.

The work here extends the Met Office's 24 ensemble members to 93. It is found that the extra members do develop a significant degree of linear independence, they increase the ensemble spread (although with caveats to do with non-Gaussianity), they reduce sampling error in many statistical quantities (namely variances, correlations, and length-scales), and improve the effective spatial resolution of the ensemble. The extra members though do not improve the probabilistic rain rate forecasts.

It is assumed that the 93-member ensemble approximates the error-free statistics, which is a practical assumption, but the data suggests that this number of members is ultimately not enough to justify this assumption, and therefore more ensembles are likely required for such convective-scale systems to further reduce sampling errors, especially for ensemble data assimilation purposes.

1 Introduction

Many operational centres use ensemble prediction systems (EPSs) to enhance the value of deterministic forecasts. An EPS allows a subset of possible alternative forecast outcomes to be assessed, and for aspects of the probability density function (PDF) of forecast uncertainty (e.g. its spread) to be estimated. These activities are useful for purposes such as forecast evaluation and data assimilation (DA). Only a relatively small number of ensemble members is affordable though and it is well known that this

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

can lead to shortcomings in sampling the probability density function (PDF) of forecast errors, especially in high-dimensional systems, like those used in numerical weather prediction.

Even when the PDF of forecast errors is truly Gaussian – in which case it is determined by its mean and covariance matrix – the *sampled* forecast error covariance matrix can represent forecast uncertainty along only a small number of directions in phase space, see e.g. Houtekamer and Zhang (2016). This issue can lead to the ensemble underestimating the true spread and it can introduce noise in the sample covariances, which is particularly evident between distant points. The ensemble data assimilation community is acutely aware of this problem, see e.g. Houtekamer and Mitchell (1998); Hamill et al. (2001); Ehrendorfer (2007); Houtekamer and Zhang (2016); Bannister (2017), where methods of mitigating sampling error have been developed such as localisation, ensemble inflation, and hybridisation. For some applications of ensemble prediction, having more than a few tens of members does not provide added benefit (Talagrand et al., 1997; Houtekamer et al., 2014), but this is not true in general. Houtekamer and Mitchell (2005) and Kondo and Miyoshi (2016) in particular argue that at least 10000 ensemble members would be needed in ensemble DA to avoid using the mitigation techniques mentioned above, but this is currently an impractical proposition for operational purposes. There are additional problems at convective-scales, where forecast errors are expected to have a significant degree of non-Gaussianity. At small horizontal scales the Rossby number is not small and the vertical and horizontal scales of motion are comparable, which means that the equations describing the evolution of the flow contain significant non-linear terms (e.g. the velocity advection terms) (Zhang, 2005). A small ensemble is therefore also likely to fail to capture adequately non-Gaussian effects such as multi-modality of PDFs. Large ensembles would also be required to capture rare events, but this aspect is not the focus of our study.

Despite these problems with sampling error in practical ensemble sizes, operational weather centres are making increasing use of convection-permitting EPSs (e.g. with the AROME model (Seity et al., 2011; Bouttier et al., 2016), the COSMO model (Gebhardt et al., 2011; Ben Bouallègue and Theis, 2014; Harnisch and Keil, 2015; Bick et al., 2016; Schraff et al., 2016), the Met UM (Bowler et al., 2008; Golding et al., 2014; Tennant, 2015), and the WRF model (Schwartz et al., 2014; Luo and Chen, 2015; Schwartz et al., 2015)[1]) with the aim of producing skillful forecasts, including forecasts of convective precipitation whose predictability can vary with time. For the reasons mentioned above, it is useful to have a range of diagnostics available to help determine how many ensemble members are required to provide a sufficiently accurate characterization of a forecast error PDF for forecasting (including nowcasting) and data assimilation purposes.

1.1 Sensitivity to ensemble size in large-scale systems

The problem of determining the forecast sensitivity to ensemble size has been investigated in large-scale systems for a number of years. Buizza and Palmer (1998) considered ensembles of 2, 4, 8, 16 and 32 members in the ECMWF system and showed that the skill of the 500 hPa geopotential height forecast increased with ensemble size, but depended on the specific forecast error norm used in a given verification method. In a related ECMWF study (Buizza et al., 1998) the benefits achieved by increasing the ensemble size versus increasing the model's resolution were studied. They considered ensembles of 32, 50

[1] AROME = Application de la Recherche à l'Opérationnel à Méso-Echelle, COSMO = Consortium for Small-scale MOdeling, UM = Unified Model, WRF = Weather Research and Forecasting model.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

and 128 members of different resolutions over 14 case studies. They showed that although increases in both ensemble size and resolution are beneficial, larger ensembles achieve a better skill. In particular they found better probability predictions of temperature and precipitation, as measured by the Brier score. Higher resolution forecasts did increase the ensemble spread, but insufficiently. Later Mullen and Buizza (2002) used the ECMWF's EPS to show that increasing only the resolution does provide clear benefits in precipitation forecast skill of lighter rain, but increasing the number of members provides better value for forecasts of heavier rain (as estimated from a simple cost-loss model). More recently Bonavita et al. (2011) used the ECMWF's ensemble of data assimilations (EDA) to show that the sample forecast error standard deviation of the vorticity fields calculated with 10 and 50 member ensembles are highly correlated (between 80% and 95%), despite the ensembles having a larger sampling noise at smaller scales. In order to reduce sampling noise, forecast error variances need to be cleaned up, which was done in spectral space using a filter with a truncation wave number estimated from climatological statistics. Forecast error standard deviations estimated with larger ensembles tend to need less filtering and therefore tend to have a higher effective spatial resolution, leading to better forecast scores.

1.2 Sensitivity to ensemble size in convective-scale systems

The conclusions of the above studies do not necessarily apply to EPSs using a convection-permitting ensemble, given the highly varying nature of predictability at convective-scales and their potential deviation from Gaussianity. For these reasons, work has also been done to study the effect of changing the number of ensemble members in convection-permitting model forecasts and some important studies are mentioned below.

Tong and Xue (2005) performed a synthetic observation study, assimilating perfectly modelled synthetic radar Doppler and reflectivity observations with an EnKF into a 2 km grid-length configuration of the ARPS[2] model. Although they used 100 members, they commented that as few as 40 members were enough to produce good analyses. Clark et al. (2011) used a 4 km grid-length configuration of WRF over the central United States. They took between 1 and 17 members over multiple cases to show that the area under the ROC curve for 6-h accumulated precipitation at various thresholds increased with increasing ensemble size for all considered scales. They found that there was little impact on the ROC area by increasing the ensemble size above 9 members, although they postulated that at least 60 members would be needed to bring sampling error and under dispersiveness down to acceptable values. They also argued that more members are required for skillful forecasts of rare events or in low-predictability regimes. Bouallegue et al. (2013) found an improvement in the reliability and resolution of precipitation ensemble forecasts by increasing the number of members from 20 to 60 in a 2.8 km grid-length version of the COSMO-DE model. Ménétrier et al. (2014) discussed the characteristics of the forecast error variances and correlation length scales for small (6 member) and large (84 member) ensembles of forecasts using the convection-permitting AROME model (2.5 km grid-length) over France, and focused on a case study characterized by strong convection. They investigated the effects of sampling errors in the small and large ensembles by comparing the forecast error standard deviations of near surface specific humidity derived from each. The small ensemble showed larger variability at small scales in particular, due to larger sampling noise. They calculated anomaly correlations between variance maps computed from the two ensembles for a range of quantities

[2] Advanced Regional Prediction System.

and found values between ∼60% and 80%. They also showed that the correlation length-scales for a range of quantities are, on average, shortened as the number of ensemble members is increased (which is consistent with the need for less localisation for larger ensembles). In the case (at least) of specific humidity fields though, their study found that this effect is dominated by a decrease in the number of instances of long correlation length-scales in their large ensemble compared to their small ensemble. The number of instances of small correlation length-scales in their large ensemble though was found actually to increase (their Fig. 13), thus complicating the effect that sampling error has on the length-scales. Schwartz et al. (2014) studied ensemble sizes of 5, 10, 20, 30, 40, and 50 members in a 3 km grid-length version of WRF over the central United States. They found that there were sometimes clear improvements in reliability, fractions skill score, and area under the ROC curve for precipitation forecasts. This was especially true for low precipitation thresholds, higher probability events, and longer forecasts. Their conclusion though was that as few as 20 to 30 members showed similar skill to the full 50 members in many respects. Harnisch and Keil (2015) found increases in Brier skill score for precipitation forecasts, and decreases in the CRPS for 10 m winds, and in the average forecast-minus-observation departures of various quantities, by increasing the number of members from 20 to 60 in a 2.8 km grid-length version of the COSMO-DE model. These benefits were found to be significantly larger than using various ensemble inflation methods, or the introduction of a model error scheme.

The overall conclusion from these studies is that the ensembles give improved forecast skill as the ensemble size is increased, but a judgment of the number of ensemble members required to achieve a particular goal depends on the model and on the application. Running high-resolution forecasts operationally remains an expensive activity, and so any studies indicating the degree of sensitivity of a range of diagnostics to ensemble size is valuable, especially when applied to models and outputs that have not been studied in this way, such as to convective-scale model forecasts of rainfall rate as is done in this paper.

1.3 Aims and scope of this paper

Throughout this work the authors had resources to generate a large ensemble for only a single convective-scale case study. This paper is intended to do four things: (i) highlight some issues around sampling error in convective-scale systems, (ii) document a means of generating a larger ensemble from an existing small ensemble, (iii) develop a number of potentially informative diagnostics, and (iv) test the diagnostics for the large ensemble. Due to the availability of only one case study, the results are therefore not intended to provide a definitive answer to the number of ensemble members required in the system studied, as to do this, more members, and more realisations would be required. It is hoped though that the methodology, the choice of diagnostics, and how they are interpreted will be useful to researchers and developers who do have the resources to use the tools to their full potential, especially those who are planning new or extended EPSs operationally.

A central assumption made in this work is that the large ensemble is sufficient to neglect sampling errors. Sub-samping from this large ensemble is then done to see how smaller ensembles (of varying size, now assumed to have sampling errors) reproduce aspects of the full ensemble. The methods are illustrated with a case study with an (up to) 93 member convection-permitting forecast ensemble based on a 1.5 km grid-length version of the Met Office's Met UM over the Southern UK. We do acknowledge that 93 members is not sufficiently large to neglect sampling errors and, as stated above, and that a single case study is not sufficient to reach a definitive conclusion, but it is beyond the scope of our project resources to generate more

members or to study further case studies. The case study is meteorologically interesting though; it is characterized by multiple rain bands generated by a cold front passing over the model's domain. This paper attempts to use and develop methods to help answer the following kinds of questions that arise in ensemble studies.

- How can linearly independent extra members be generated from an existing ensemble?
- How can the ensemble size impact probabilistic forecasts of rainfall?
- How can the ensemble size affect how the kinetic energy spectrum is resolved?
- How can the ensemble size affect estimates of (co-)variability of thermodynamic and moisture fields?
- Is the ensemble used in a particular application large enough to neglect sampling error?

The structure of this paper is as follows. Section 2 is a description of the case study, Sect. 3 explains how the 93-member ensemble is created from the operational 24-member ensemble and examines how linearly independent the extra members are. The remaining sections describe how sensitive various diagnostics are to ensemble size. Section 4 looks at ensemble means and spreads, Sect. 5 looks at probabilistic aspects, Sect. 6 looks at kinetic energy spectra, and Sect. 7 looks at aspects of the forecast error (co)variances. Finally, Sect. 8 discusses the main conclusions.

2 The 20th September 2011 case study

This case study is of interest to the DIAMET[3] project. It comprises multiple cloud bands (labelled "1", "2", and "3" in Fig. 1) over Southern UK and it was intensively observed with a field campaign (Vaughan et al., 2015). The case is also studied by Baker et al. (2014) using a 24-member Ensemble Transform Kalman Filter (ETKF) ensemble mentioned in Sect. 3.4. The case is characterised by an eastward moving cold front over the southern UK as shown in Fig. 3 of Baker et al. (2014). Rainfall maps constructed from the UK's network of radar instruments are shown in Fig. 1 from 13Z (panel a) to 18Z (panel f). At 13Z there are three rain bands associated with the front, but only band 1 is within the domain at this time. Band 2 enters the domain at 14Z, and becomes very clear at 15Z and 16Z. Rain band 3 has just started to appear in the boundary of the domain at 16Z, and all three bands are within the domain at 17Z and 18Z. Bands 1 and 2 have started to merge at 18Z. Parts of bands 2 and 3 are extremely thin at 17Z and 18Z which makes them difficult features to expect the model to capture. More details about this case are given by Baker et al. (2014), and here we focus mainly on the ensemble at 15Z.

[3]DIAbatic influences on Mesoscale structures in ExTratopical storms.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

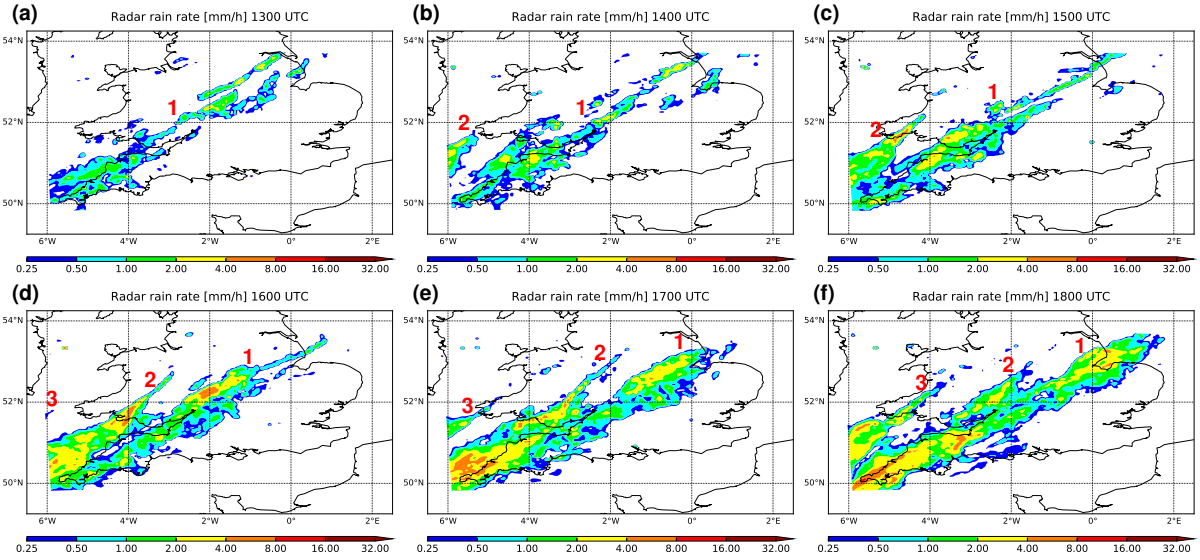

Figure 1. Radar rainfall rates for 20th September 2011 for (a) 13Z, (b) 14Z, (c) 15Z, (d) 16Z, (e) 17Z, and (f) 18Z. The three rain bands discussed in the text are labeled.

3 Generating more members from an existing ensemble

3.1 The meteorological model and the existing ensemble prediction system

The model used for this study is a 1.5 km grid-length version of the Met Office's Unified Model for a region over the southern UK (SUK-1.5[4]). This model is no longer operational, but was used during the London 2012 Olympics Golding et al. (2014); Ballard et al. (2016). The model's initial conditions (ICs) and lateral boundary conditions (LBCs) were produced using a set of nested forecasting systems. The starting point was the Met Office's global model, which produced the ICs and LBCs for an intermediate resolution regional model (the North Atlantic and Europe [NAE] domain), which in turn produced ICs and LBCs for the convective-scale SUK-1.5 model.

The existing EPS used for this study was the 24-member MOGREPS-G system (the global configuration of the Met Office Global and Regional Ensemble Prediction System (Bowler et al., 2008)). This was run with the global model, whose ensemble members were updated on a 6-h cycling timescale using an ensemble transform Kalman filter (ETKF) – for perturbations – and a 4DVar analysis – for the control – which the analysis perturbations were centred about (Migliorini et al., 2011; Caron, 2013). The standard 24 MOGREPS-G system comprised 23 perturbations and one control member, here denoted (23+1).

[4]SUK-1.5 was the designation of this model in Bannister et al. (2011), although the model is also known as the Nowcasting Demonstration Project (NDP) by the Met Office.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

3.2 Generating the 93 convective-scale ensemble members

Members of a forecast ensemble may be constructed in different ways. Perhaps the most justifiable method is to generate an analysis ensemble using an Ensemble Kalman Filter (EnKF), or ETKF (Ehrendorfer, 2007; Houtekamer and Zhang, 2016), and then to propagate this ensemble forward in time. As long as the EnKF is set-up in an optimal fashion, that the ensemble is large enough to neglect sampling errors, and that errors are close to Gaussian, such an ensemble should have the correct spread and contain the correct correlations between variables, which reflect the true uncertainty of the system given the available data. This method relies on the existence of a forecast ensemble in the first place, so it cannot be used directly to generate an ensemble from scratch.

Some methods are more suitable to generate an ensemble from scratch. The breeding method (Toth and Kalnay, 1997; Corazza et al., 2003; Wang and Bishop, 2003) starts with random perturbations to a state. It then uses the numerical model to propagate the perturbed and unperturbed states over a defined time interval, and then scales-down the propagated perturbations to a specified size. This process is repeated until the states (which become the ensemble members) have lost memory of the initial random perturbations. The states generated are called bred vectors and recognise the non-linearity of the system. The singular vector method (Buizza and Palmer, 1995) on the other hand finds the perturbations to a linearised system that grow the fastest over a defined time interval. Each perturbation is constrained to have a specified covariance amongst its elements at the start time (typically the analysis covariance is used, in which case the states are called Hessian singular vectors). Singular vectors though do not account for non-linearity of the model. The system simulation approach (Houtekamer et al., 1996) builds ensembles by performing multiple forecast and DA cycles for a set of initially random states, while attempting to vary all known uncertainties. Each state is fed through these cycles with perturbed observations, perturbed aspects of the model (e.g. the parametrisation schemes) and perturbed ancillary fields (e.g. sea surface temperatures). This is repeated until the ensemble statistics become stable. The ECMWF uses a similar approach to maintain its ensemble system, e.g. Bonavita et al. (2012). The initial members may be sampled from the climatological background error covariance matrix (as in e.g. Raynaud and Bouttier (2016)). Alternatively, a set of ensemble members that sample a climatological PDF may be taken from one long (e.g. multi-year) model run. In this method, the ensemble is built by extracting states over a specific season of this run. This is the approach used by Miyoshi et al. (2014) for their very large ensemble.

In this study we already have a 24-member ensemble and we develop the ability to generate additional members. The extra members are generated by negating existing perturbations. This method is conceptually simple, but is complicated by the nesting and sub-nesting of the various models involved. Three separate modelling systems are involved: the global model, the NAE model, and the SUK model. The procedure is illustrated in Fig. 2, and the upper case labels (A), (B), etc., in the following refer to this chart.

1. Run the (23+1)-member MOGREPS-G system (global model, 60 km grid-length, 70 levels) to generate ICs and LBCs for the NAE model. The chosen start time is 12Z on 16th September 2011.
   (a) Negate the 23 perturbations to generate 23 more (46+1).
   (b) Spin-up the global model for 48 hours (6 hour DA cycling) with the (46+1) members (B).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

(c) Negate the 46 perturbations to generate 46 more (92+1).

(d) Run-in the global model for a further 54 hours (6 hour DA cycling) with the (92+1) members (C).

2. Downscale the (92+1) global members from 1(d) to the NAE domain (18 km grid-length, 70 levels). The LBCs needed for the NAE model are also generated from step 1(d) during the global model integrations.

(a) Run the NAE model for 12 hours with the (92+1) members (D).

3. Downscale the (92+1) NAE members to the convective-scale grid (SUK domain, 1.5 km grid-length, 70 levels) to generate convective-scale initial conditions. The LBCs needed for the SUK model are also generated from step 2.

(a) Run the convective-scale model for 11 hours (no DA) with the (92+1) members (E).

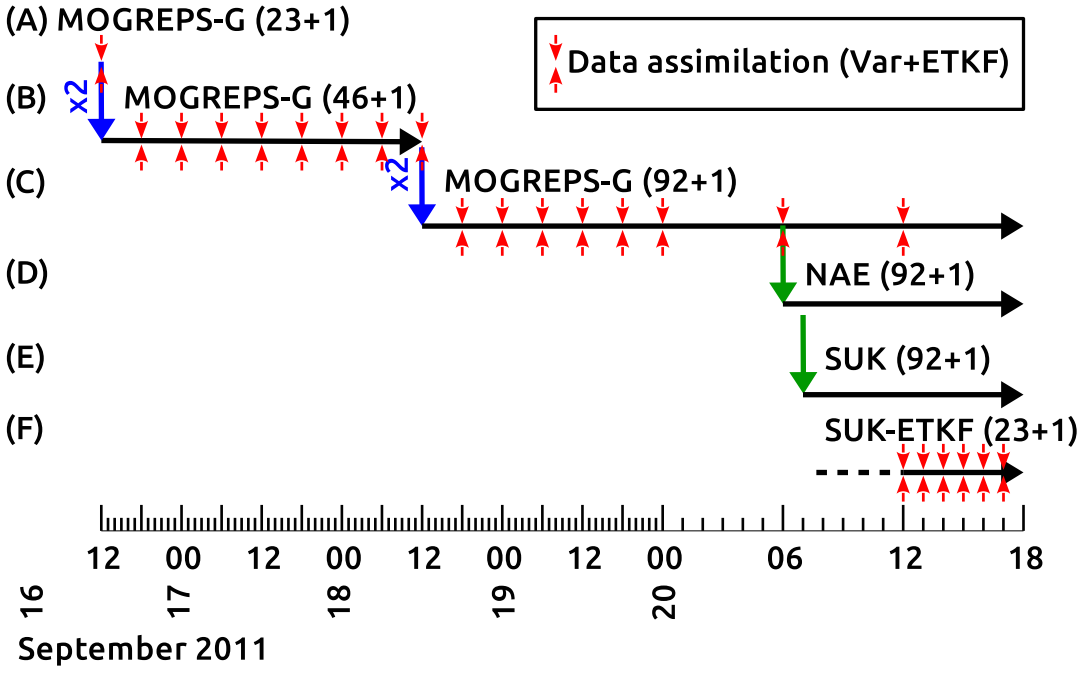

Figure 2. Outline of how the (92+1) member convective-scale ensemble forecasts were created. Each horizontal line represents a different run of an ensemble system. (A) global (23+1) archived analysis members for 12Z 16th September 2011; (B) global 48 h run of the (46+1) member ensemble; (C) global 48 h run of the (92+1) member ensemble; (D) regional (NAE) 12 h run of the (92+1) member ensemble; (E) regional (SUK) 12 h run of the (92+1) member ensemble; (F) SUK run of the (23+1) member ensemble with DA (from Baker et al. (2014)). The red arrows indicate that DA is performed (ETKF with re-centring on the Var control state), the blue down-arrows labeled "x2" indicate a doubling of the number of perturbations by negation, and the green down-arrows indicate downscaling of the members to a finer grid.

The creation of the extra perturbations by negating the existing ones (as in steps 1(a) and 1(c)), can be continued to generate exponentially more ensemble members as required, and as resources allow.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

3.3 Linear independence of the extra members

The extra perturbations are not immediately linearly independent, but it is reasonable to assume that independence will develop during the run of the non-linear models and that this will happen adequately during the 36 hours from the start of the (92+1)-member MOGREPS-G run to the start of their downscaling to the NAE, and thereafter to the SUK model. Gilmour et al. (2001) showed that similar positive/negative perturbations in global models usually take no more than 48 hours (and often less than 24 hours) to show significant non-linearity. This is expected to happen much more quickly in a convective-scale model. Given a set of ensemble perturbations at a particular time the degree of independence can be checked by attempting to construct a set of orthogonal vectors. Considering each perturbation in turn, components of perturbations already considered are systematically subtracted. The size of the residual is then a measure of the linear independence of the perturbation considered with respect to the previous perturbations. Let $\mathbf{x}'_i$ be the perturbation of member i (with respect to the control member), and let $\hat{\mathbf{x}}'_i$ be the ortho-normalised vector satisfying ${\hat{\mathbf{x}}'_i}^{\mathrm{T}}\hat{\mathbf{x}}'_j = \delta_{ij}$. $\hat{\mathbf{x}}'_i$ can be constructed from the Gram-Schmidt procedure:

$$\hat{\mathbf{x}}'_i = \frac{1}{\hat{N}_i}\left(\frac{\mathbf{x}'_i}{N_i} - \sum_{j=1}^{i-1}\alpha_{ij}\hat{\mathbf{x}}'_j\right), \tag{1}$$

where $N_i^2 = {\mathbf{x}'_i}^{\mathrm{T}}\mathbf{x}'_i$ and $\hat{N}_i^2 = {\boldsymbol{\psi}'_i}^{\mathrm{T}}\boldsymbol{\psi}'_i$ (where $\boldsymbol{\psi}'_i$ is the state in brackets in (1)) are for normalisation, and α_{ij} is chosen for orthogonality, $\alpha_{ij} = {\mathbf{x}'_i}^{\mathrm{T}}\hat{\mathbf{x}}'_j/N_i$. If $\hat{N}_i$ is found to be unity then the perturbation $\mathbf{x}'_i$ is already perfectly orthogonal (linearly independent) to the previous perturbations, and if $\hat{N}_i$ is found to be zero then it does not contain a new direction in phase space[5].

Figure 3 plots the value of $\hat{N}_i$ as a function of ensemble perturbation member number and time of day, considering separately a single level of (a) temperature and a single level of (b) specific humidity. Even though the $\boldsymbol{\psi}'_i$ should represent the whole state, rather than specific variables and levels, we assume that it is sufficient to look at this issue in relevant subspaces (a single variable/level has a subspace size of $360 \times 288 = 103680$). In Fig. 3 the first perturbation at the bottom of each column of bars is always unity by construction, and so acts to set the scale. Although $\hat{N}_i$ broadly decreases with perturbation number for the temperature and specific humidity subspaces, the values remain significant, and no perturbation has $\hat{N}_i \approx 0$. This is a demonstration that each respective member has developed some unique information to add to the ensemble as a whole.

3.4 The ensembles used in this paper

Although the discussion in Sect. 3.2 emphasises the large ensemble, a number of ensembles are used in this paper. These are summarised as follows:

- The 'large ensemble', $N = 93$, uses all (92+1) members.

[5]Strictly the orthogonality would be better defined in a norm that respects the PDF of forecast errors, but we use the Euclidean norm for simplicity. The degree of linear independence can differ significantly between these norms if the covariance of the forecast error PDF has a high condition number.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

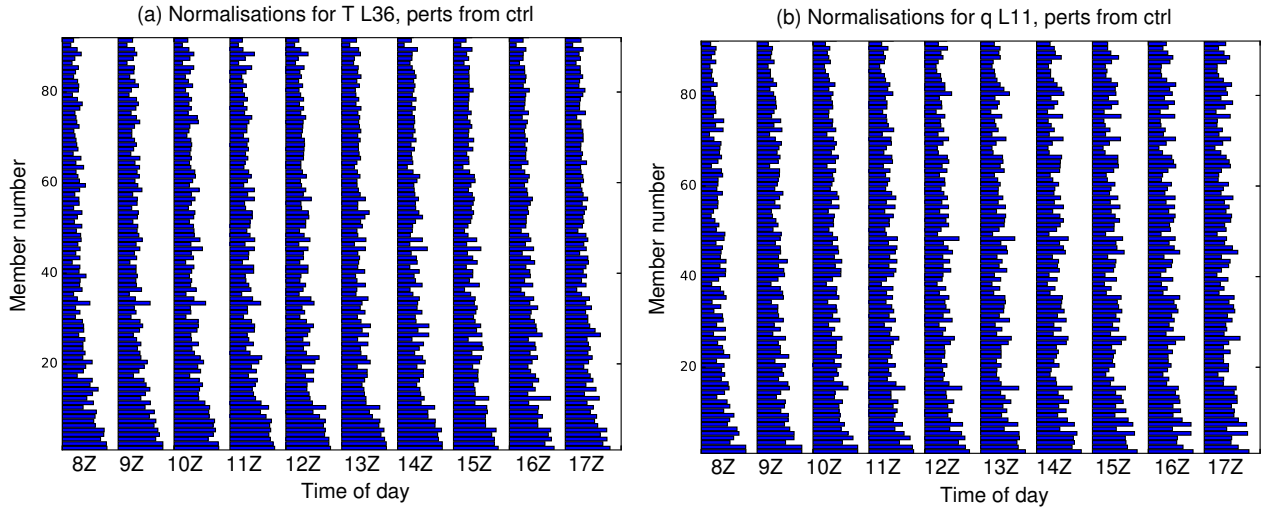

Figure 3. Values of $\hat{N}_i$ for the large ensemble for (a) temperature perturbations at model level 36, and (b) specific humidity perturbations at level 11 for different times of the day on 20th September 2011. $\hat{N}_i$ indicates the degree of independence of each perturbation over previous perturbations (1 means fully independent and 0 means not independent at all). The first member (the lowest bar in each column) represents $\hat{N}_i = 1$.

- The 'intermediate ensemble', $N = 47$, has (46+1) members. These are randomly sub-sampled from the large ensemble a number of times, although the control member is always present. There are $\sim 4 \times 10^{26}$ such unique sub-samples. Depending on the experiment, we either take one sub-sample at random, or a reasonably large number (1000).
- The 'small ensemble', $N = 24$, has (23+1) members. These are sub-sampled from the large ensemble in the same way as for the intermediate ensemble. There are $\sim 3 \times 10^{21}$ unique sub-samples, but we again take either one or 1000 at random.
- The 'ETKF ensemble', $N = 24$, is a single ensemble comprising the (23+1) member ensemble used in a previous study, where a convective-scale ETKF is used to update the state over an hourly cycle (Baker et al., 2014).

In the case of the sub-sampled ensembles, whether just one sub-sample is taken or 1000 depends on the context, and the control member is always included in sub-samples. How many sub-samples used is specified in each figure caption.

4 Sensitivity of means and spreads of forecasts ensembles to ensemble size

4.1 Impact on the ensemble mean and the rain banding

Figure 4 shows various representations of the 15Z (T+8) rainfall rates. Panels (a)-(d) comprise four members from the large ensemble, which can be compared to the rainfall rate patterns observed by the radar, panel (e). All of the 93 members predict

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

a rain band, but none gets all aspects of the measured rainfall correct (i.e. position, orientation, size, and strength). The vast majority of members do not show multiple rain bands, but four do have two clear rain bands, and an example is shown in panel (b) (note that rain band 3 is outside of the SUK domain at 15Z). The rain rates of the individual members are comparable to the radar.

The remaining panels of Fig. 4 are ensemble means for the ensembles discussed in Sect. 3.4. These are for the large ensemble (panel f), the intermediate ensemble (a single sub-ensemble of (46+1) members, panel g), the small ensemble (a single sub-ensemble of (23+1) members, panel h), and the ETKF ensemble (i). As expected the ensemble means have patterns that are broader and less intense than the individual members. This is due to the averaging of members at slightly different locations, and results in a mean field that is not very useful (Ancell, 2013; Hollan and Ancell, 2015). There is arguably marginally more noise in the ensemble means for smaller ensemble sizes. This indicates qualitatively that to go from (23+1) to (92+1) members does not achieve significant improvement in the ensemble mean precipitation forecasts for this case. This is consistent with the lack of sensitivity of ensemble mean forecast skill to ensemble size experienced by Buizza and Palmer (1998), and also discussed in Leith (1974).

These results are also compared with the ensemble mean rainfall from the (23+1) member ensemble of Baker et al. (2014), Fig. 4i. This is the ETKF ensemble (row F of Fig. 2), and is derived from a shorter forecast (T+3) initialised from an ETKF and 3DVar at 12Z. This ETKF ensemble also does not show two rain bands (see Sect. 4 of Baker et al. (2014)) and the band that is forecast is also fairly comparable to that in the large ensemble, albeit with more intense rain and better alignment with the observed rain band.

4.2 Impact on the relative standard deviation

Figure 5 shows the relative standard deviation (the ratio of the sample standard deviation to the sample mean where the mean rain rate is greater than 0.01 mm/h) for each of the ensembles. Each map shows largest values close to the edge of its respective rain band in Fig. 4 and is well aligned with each band's main axis. This shows that the main rain band tends to have approximately the same orientation but slightly different locations amongst the members of all ensembles. The relative standard deviation for the ETKF ensemble shows more small-scale variability than the others and we do not know whether this is a consequence of the application of the ETKF at the convective-scale, or just the fact that this is a shorter forecast. The large ensemble presents larger values, and a larger width of the standard deviation strip and is a consequence of the wider variability of the centres of the rain bands amongst its ensemble members. The qualitative similarity of the maps for the extended ensembles and the small ensemble in Figs. 4 and 5 suggests that the method used to generate extra members described in Sect. 3.2 produces a physically reasonable ensemble.

To assess how different the ensembles are at capturing higher moments of the 'true' rainfall rate distribution, Fig. 6 shows aspects of the sampled distributions of domain and ensemble averaged rainfall for the large (red) and small (dashed blue) ensembles at progressively longer lead times. In this Fig. the horizontal line within each box represents the median value, the top and bottom edges of each box represent the interquartile ranges (25% change in cumulative distribution function from the median), and the whiskers denote the minimum and maximum values. This plot shows that the large and small ensembles each

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

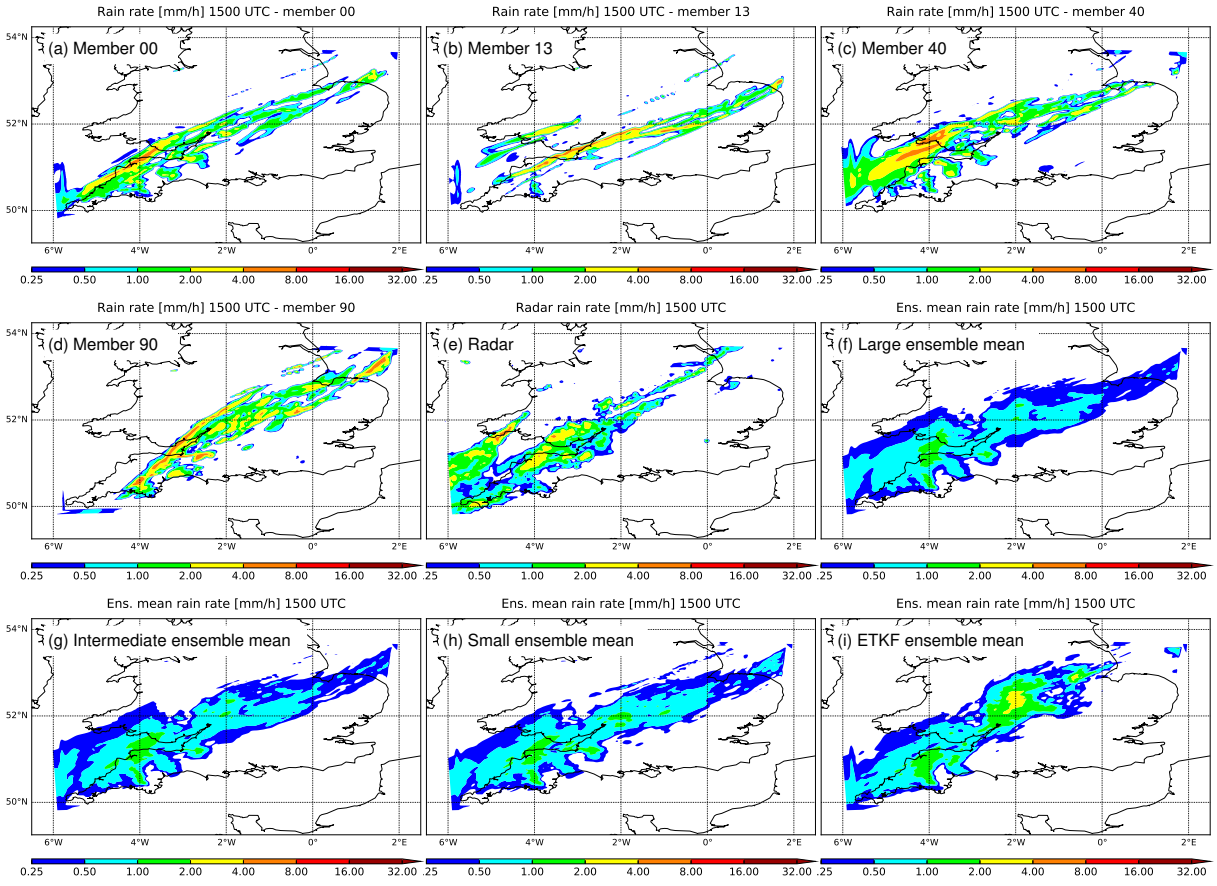

Figure 4. Rain rates for 20th September 2011 15Z. Panels (a)-(d) comprise a selection of individual ensemble members from the large ensemble, panel (e) is the radar composite, and panels (f)-(i) are ensemble means from the ensembles listed in Sect. 3.4, where the intermediate and small ensembles each represents a single sample from the large ensemble.

do a similar job in representing some aspects of the non-Gaussian rainfall distribution, with the range of the small ensemble being smaller than that of the large ensemble as expected (note the logarithmic scale of the y-axis). Although we would expect the small ensemble to have a smaller spread than the large ensemble (as found e.g. in Fig. 5), the interquartile ranges are larger in the small ensemble for five of the ten times shown, namely at 2, 4, 6, 7, and 8-hour lead times (note that for a similar interquartile range of the large and small ensembles, the dashed lines marking the edges of each small ensemble's interquartile range should lie at the centre of the thick red lines marking the edges of the large ensemble's range; note also the logarithmic scale). the 8-hour lead time corresponds to 15Z, which shows an increase in variance with more members in Fig. 5). This is an example of a case when the variance is not always a complete way of describing a non-Gaussian distribution. The variability in the mean between the large and small ensembles is insignificant compared to the spreads.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

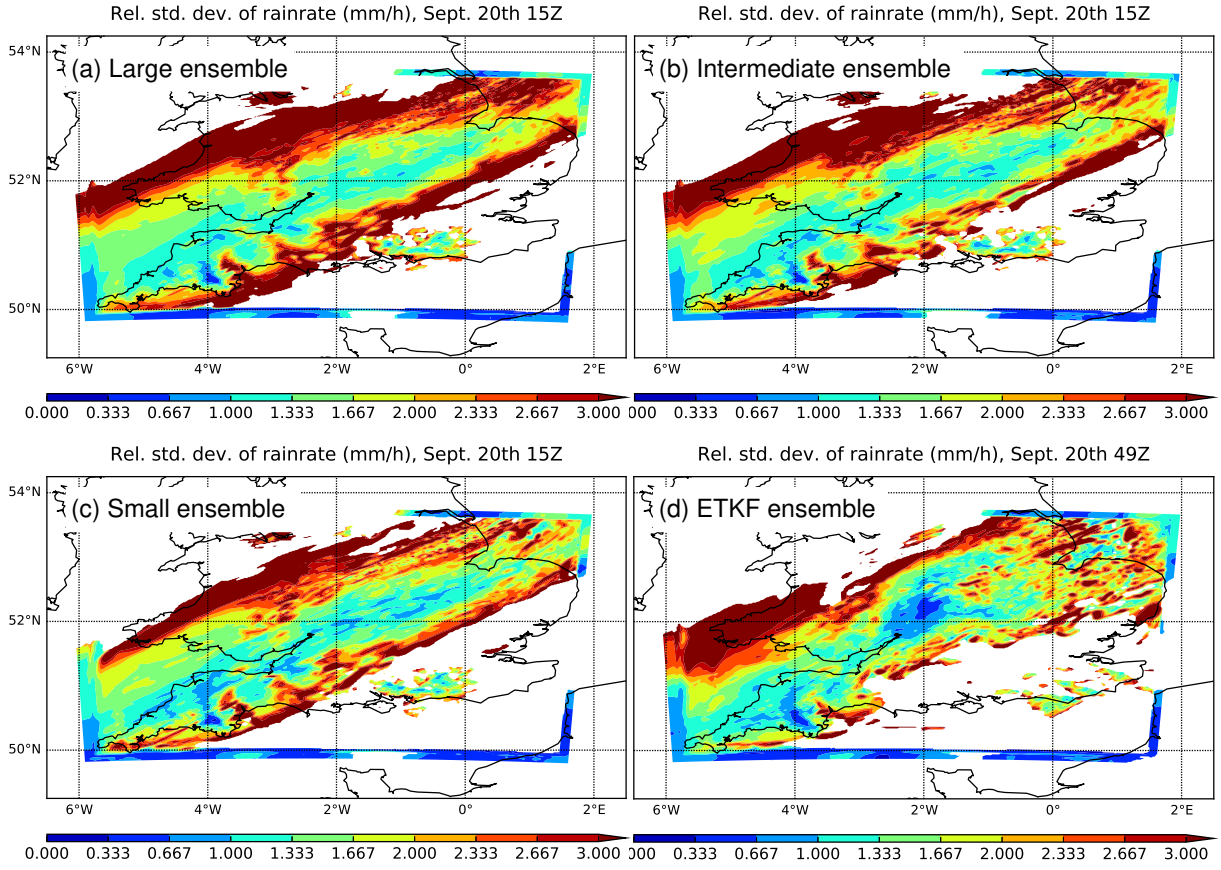

Figure 5. Relative standard deviations of rain rate for 20th September 2011 15Z for the large ensemble (a), the intermediate ensemble (b), the small ensemble (c), and the ETKF ensemble (d). The relative standard deviation rain rate is defined as the ratio between the sample standard deviation rain rate to the ensemble mean rain rate (defined where the mean rain rate is greater than 0.01 mm/h). The intermediate and small ensembles each represents a single sample from the large ensemble.

4.3 Verification of the rainfall spread

The rainfall forecast ensemble is now verified against the radar observations with rank histograms. The ensemble of rain rate forecasts (valid on 20th September 2011 at 15Z) provide the ranks at each point in the model's domain, which are populated by radar measurements that are above a specified threshold rain rate. The ranks are determined at each point according to the sorted ensemble of rain rates, where each is modified by a value drawn from the observation error distribution (Hamill, 2001). This distribution is assumed to be Gaussian with a standard deviation of 0.316 mm/h, as used in Migliorini et al. (2011). Figure 7 shows rank histograms for the small (blue), intermediate (green), and large (red) ensembles, for the thresholds of (a) 0.0 mm/h, (b) 0.02 mm/h, and (c) 0.2 mm/hr. The large ensemble histogram is reproduced in both columns to aid comparison with the smaller ensembles. A correctly spread ensemble is expected to have a flat rank histogram.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions
Open Access
EGU

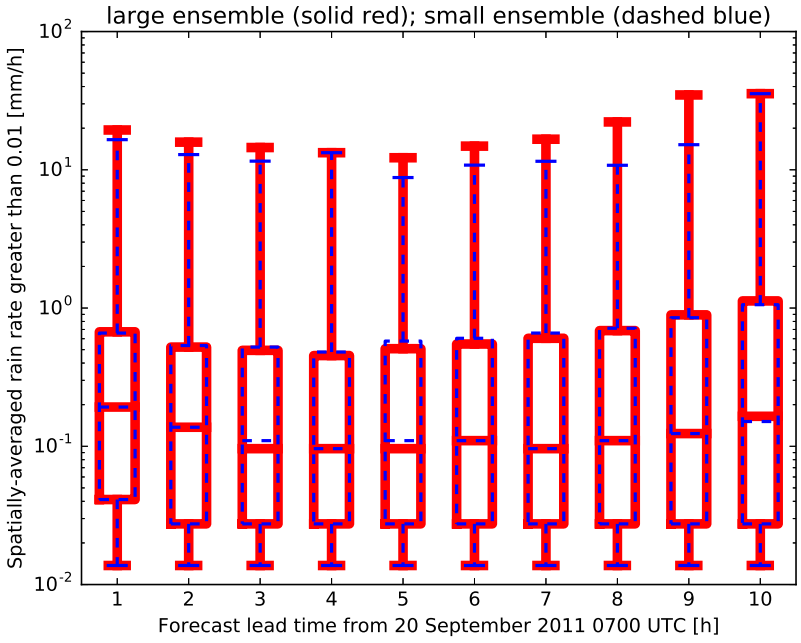

Figure 6. Domain averaged rainfall rate (mm/h) distribution as a function of forecast lead time for the large ensemble (solid red) and the small ensemble (dashed blue). The horizontal line within each box represents the median, the top and bottom edges of each box represent the interquartile ranges, and the whiskers denote the minimum and maximum values. The small ensemble represents a single sample from the large ensemble.

Focusing first on the large ensemble, the histogram for the 0.0 mm/h threshold (panel a) suggests that the ensemble is over-spread, and with a positive bias (the ensemble tends to over-estimate the observed rainfall). The picture changes after including only grid-points that have more than 0.02 mm/h rain rate (b). This histogram suggests that the ensemble is still over-spread, but to a much lesser extent than in (a). The histogram found by including only grid-points that have more than 0.2 mm/h rain rate (c) is similar to (b), but with evidence of a negative bias in rain rate. These mixed results mean that the ensemble exhibits different characteristics depending upon the intensity of the actual rainfall.

Similar conclusions follow from the respective histograms computed from the small and intermediate ensembles shown in Fig. 7. The apparent consistency of the degree of bias for different ensemble sizes is evidence that rank histograms give a robust estimate of this quantity even for small ensemble sizes. It is clear though from panel (a) that the large and intermediate ensembles have more spread than the small ensemble. Since the spread of even the small ensemble is too large, the extra spread of the larger ensembles worsens this diagnostic. This increase is due to reduced sampling error and so this is an example of a result that gives a misleading negative impact of the extra members.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

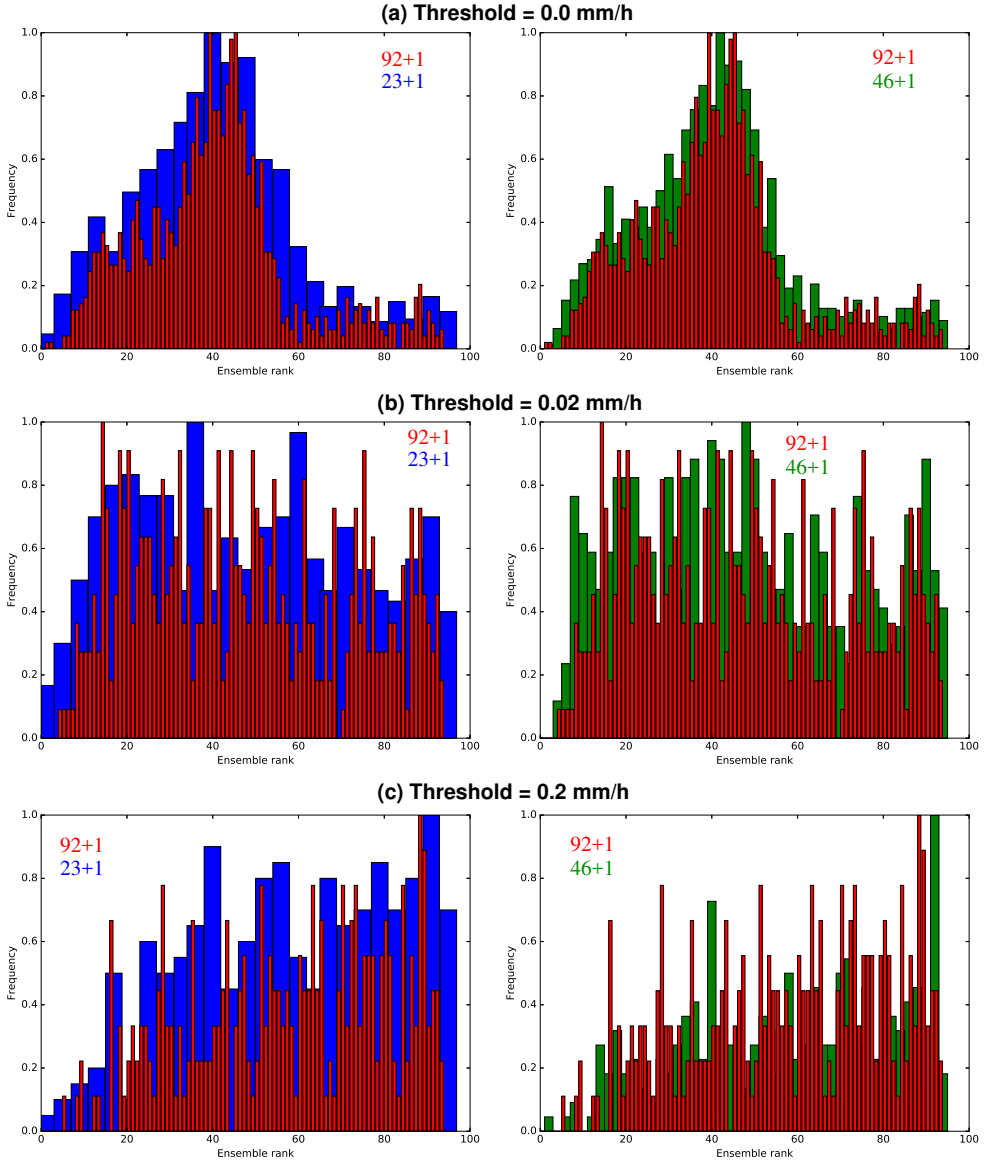

Figure 7. Rank histograms computed from the small (blue bars, left column), intermediate (green bars, right column), and large ensembles (red bars, both columns), for the thresholds of (a) 0.00 mm/h, (b) 0.02 mm/h, and (c) 0.20 mm/h. Data are for 20th September 2011 15Z. The intermediate and small ensembles each represents a single sample from the large ensemble.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

5 Probabilistic characteristics of rainfall forecasts

It is well known that an ensemble of forecasts provides a means of estimating the probability of certain events happening according to region, and an available ensemble that has many more members than the operational ensemble is a resource to study whether probabilistic diagnostics can be improved by increasing the number of members.

Figure 8 shows the probability of rain for the large and small ensembles for three rain rate thresholds, 0.02, 0.2 and 2 mm/h. The regions of high probability do not change considerably between the large and small ensembles for all thresholds, although they do of course reduce with threshold. The main differences between the large and small ensembles are that the probability maps for the large ensemble are smoother and have slightly smaller values than for the small ensemble (a similar conclusion follows for other thresholds and lead times, not shown). As with previous results, these effects can be attributed in principle to the greater variation in positioning of the rainfall between members of the large ensemble than the small ensemble, rather than fundamentally different behaviour of the extra members.

Comparing these probability plots to Fig. 1c allows us to assess whether the ensemble predicts a possibility of rain at all areas where there was rain measured by the radar. This is broadly true at the position of band 1 (but only in patches for the 2 mm/h threshold), but there is a low probability of rain at the position of rain band 2 ($\lesssim$20% chance of rain at the 0.02 mm/h threshold, and virtually zero chance of rain at the 2 mm/h threshold). It is interesting to note that even though the probability maps of any rain (0 mm/h threshold, not shown) are virtually indistinguishable from the 0.02 mm/h maps in panels (a) and (b), the rank histograms for 0 mm/h and 0.02 mm/h in Fig. 7 are strikingly different.

6 Sensitivity of dynamical aspects to ensemble size

It is important to check how this ensemble provides a dynamical description of the flow. An effective way to investigate this issue is to look at the kinetic energy (KE) spectrum as represented by the large and small ensemble systems. This is done using the following procedure (broadly following Skamarock (2004), and omitting field position and time arguments for brevity).

1. Compute the weighted wind components (for each position, time, and ensemble member) as follows: $U_i = u_i\sqrt{\rho_i \Delta_i}$, $V_i = v_i\sqrt{\rho_i \Delta_i}$, $W_i = w_i\sqrt{\rho_i \Delta_i}$, where u_i, v_i, and w_i are respectively the zonal, meridional, and vertical wind components, ρ_i is the density, and Δ_i is the grid-box thickness ($\left(U_i^2 + V_i^2 + W_i^2\right)/2$ is then proportional to the KE of each grid box).

2. Modify each of U_i, V_i, and W_i to remove the linear trend in the W-E direction. This is done separately for each ensemble member, time, vertical level, and latitude row, and is done to make the field periodic along the horizontal grid lines in preparation for step 3 (Errico, 1985).

3. Perform a discrete Fourier transform of U_i, V_i, and W_i along the longitudinal direction only, for each ensemble member i. This produces fields $\bar{U}_i$, $\bar{V}_i$, and $\bar{W}_i$, which are each a function of longitudinal wavenumber k, latitude, vertical level, and time.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

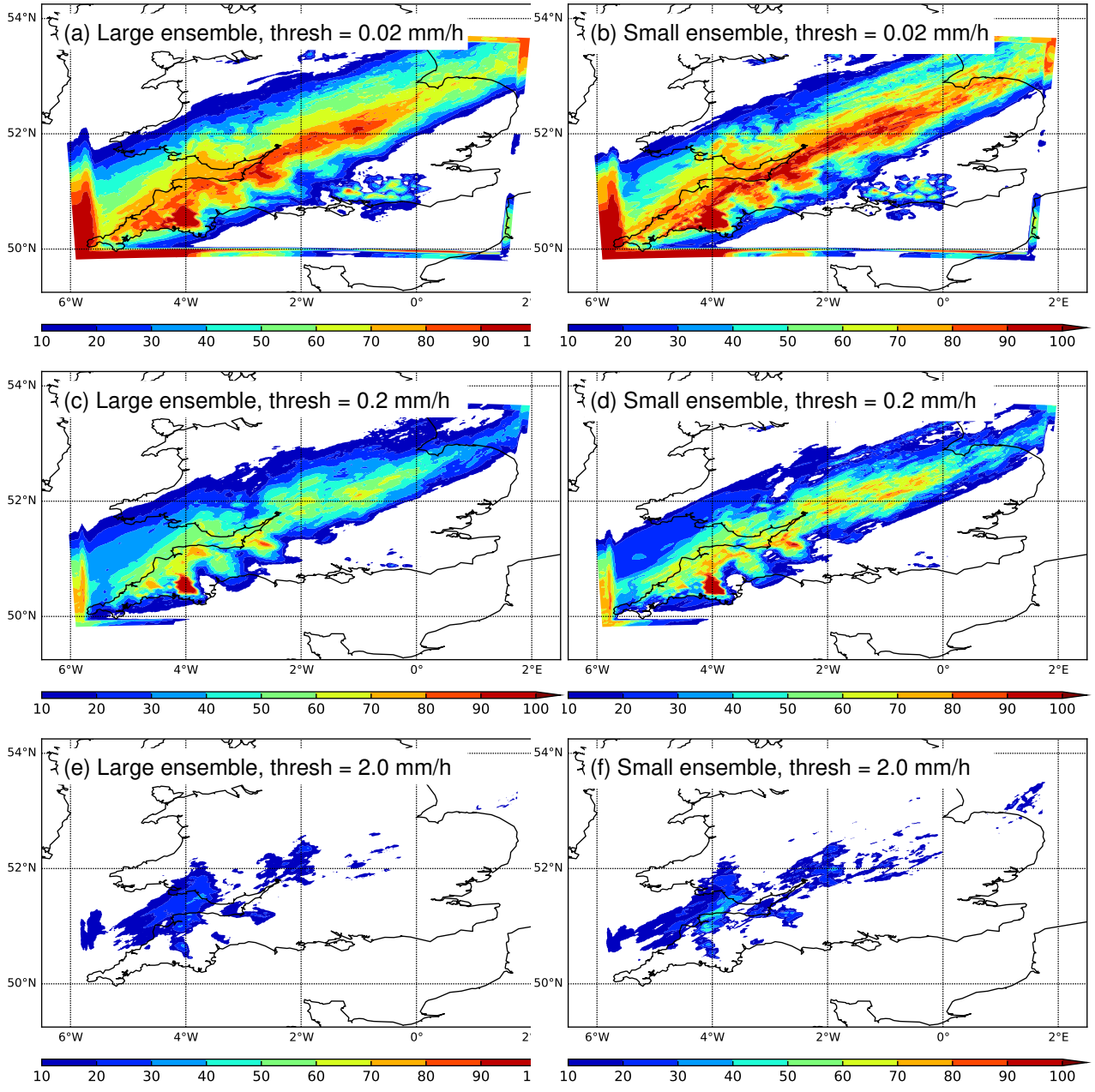

Figure 8. Probability of rain for 20th September 2011 15Z based on the percentage of ensemble members that forecast rain in each grid box above thresholds of >0.02 mm/h (top row), >0.2 mm/h (middle row), and >2.0 mm/h (bottom row). Plots are for the large (left column) and the small (right) ensemble. The intermediate and small ensembles each represents a single sample from the large ensemble.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

4. Compute the square of the modulus of these transformed fields at each wavenumber, latitude, level, and time for the KE, $\bar{E}_i = (\bar{U}_i\bar{U}_i^* + \bar{V}_i\bar{V}_i^* + \bar{W}_i\bar{W}_i^*)/2$, where * is the complex conjugate.

5. Average $\bar{E}_i$ over the domain in latitude, and over the free troposphere (between levels 30 and 51, i.e. between ~ 3 and ~ 9 km above sea level), in time (between 12 and 17Z), and over ensemble members. This results in a KE power spectrum, $\bar{E}^{(N)}(k)$, that is a function of k only, where N indicates the number of ensemble members. Furthermore, the standard deviation, $\bar{\sigma}_E^{(N)}(k)$, amongst the members is also computed.

Note that points within 10 grid-points of the lateral boundaries were omitted in all calculations.

$\bar{E}^{(93)}(k)$ is shown in Fig. 9 (red dots). This shows a characteristic upturned tail at high wavenumbers typical of spectra from limited-area models (Skamarock, 2004). The variability of the spectrum, $\bar{\sigma}_E^{(N)}(k)$ (red error bars) is also shown. From this figure, three main conclusions may be drawn.

1. The slope of the spectrum provides information about the type of turbulence. The slope is much closer to that of k^{-3} (characteristic of 2D turbulence, as expected at larger scales in the mid-latitude atmosphere) than that of $k^{-5/3}$ (characteristic of 3D turbulence, as expected at convective-scales).

2. The effective spatial resolution of the ensemble – here defined as the wavelength at which the slope of $\bar{E}^{(N)}(k)$ (in the log/log plot) starts to deviate from the best-fit slope at larger wavelengths (Skamarock, 2004) (by eye taken to be that of the k^{-3}) – is ~ 8 km ($\sim$ 5-6 grid-lengths). This is also the scale below which the uncertainty of the KE grows considerably. Scales smaller than 8 km are referred to as the unresolved scales.

3. $\bar{E}^{(N)}(k)$ has relatively little variability except at the unresolved scales.

The absence of a $k^{-5/3}$ spectrum seems puzzling as the small-scale flow is not expected to be 2D stratified like mid-latitude large scale flow. These results though do appear to be consistent with Laprise et al. (2008) who also looked at the variance of small-scale motion in nested models without small-scale driving information (provided by DA). They found that small-scale motion develops through the large-to-small-scale energy cascade over a period of a few days. Since our SUK model data used for Fig. 9 are a maximum of only ten hours old, this may explain why a small-scale (3D-like) energy spectrum is not seen. Raynaud and Bouttier (2016) also showed that perturbations downscaled into the 2.5 km grid-length AROME EPS took 9-12 hours to develop into realistic small-scale structures. This issue may be particularly relevant to our case study due to the passage of a cold front, as opposed to a case where convection is the dominant source of precipitation (the latter of which may generate 3D motion more quickly).

To quantify how the ensemble size affects the sampled variability in $\bar{E}^{(N)}(k)$, the relative standard deviation, $\bar{r}_E^{(N)}(k) = \bar{\sigma}_E^{(N)}(k)/\bar{E}^{(93)}(k)$, is computed and shown in Fig. 10a for the large (red dots), intermediate (green) and small (blue) ensembles. This is the first diagnostic shown in this paper where it is meaningful to average the standard deviations of the small and intermediate ensembles over 1000 randomly selected sub-samples from the large ensemble, rather than from just one sub-sample. The relative standard deviations are small for all ensembles for scales above ~ 8 km (the resolved scales according to the above interpretation), and grow rapidly for small scales. The standard deviations increase with the ensemble size, which

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

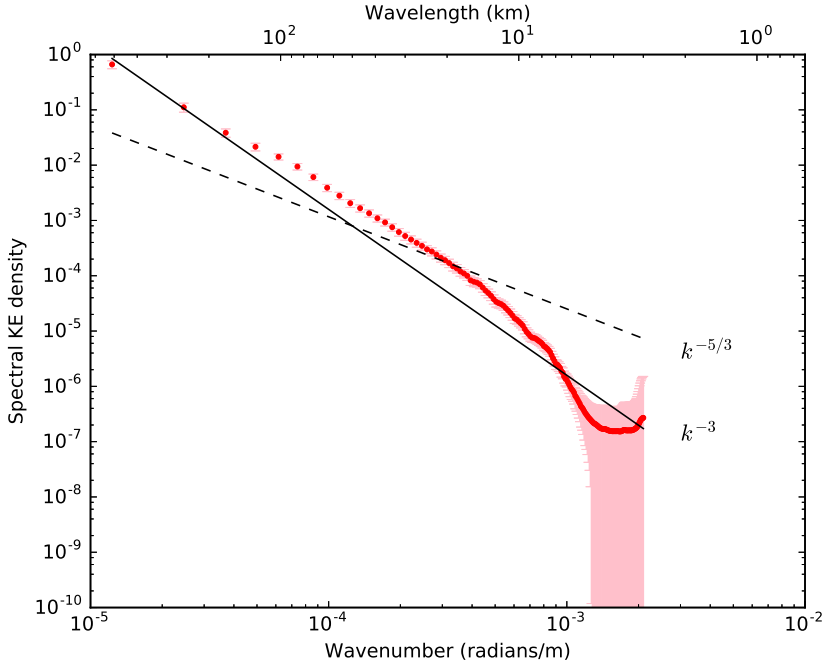

Figure 9. The large ensemble's mean kinetic energy spectrum $\bar{E}^{(93)}(k)$ (red dots), the sample standard deviation (red error bars), and the $k^{-5/3}$ and k^{-3} lines for comparison. See the text for a description of how the kinetic energy spectrum is computed.

is consistent with the expectation that smaller ensembles underestimate the standard deviation. As an indication of the larger estimated sampling errors of the small ensemble over the intermediate ensemble, the estimated RMSE in $\bar{\sigma}_E^{(47)}$ (i.e. $\bar{\sigma}_E^{(47)} - \bar{\sigma}_E^{(93)}$) averaged over resolved scales down to 8 km is 0.41 of that of $\bar{\sigma}_E^{(24)}$ (this is not visible on the scale of Fig. 10a), and averaged for the unresolved scales the fraction is 0.36. This suggests the potential degree of improvement as a consequence of nearly doubling the number of ensemble members of the original MOGREPS system.

Assuming that the value of $\bar{r}_E^{(93)}(k)$ has negligible error (and thus a proxy for the truth), the (sampling) errors $\Delta\bar{r}_E^{(24)}(k) = \bar{r}_E^{(24)}(k) - \bar{r}_E^{(93)}(k)$, and $\Delta\bar{r}_E^{(47)}(k) = \bar{r}_E^{(47)}(k) - \bar{r}_E^{(93)}(k)$ are found and plotted in Fig. 10b. The sampling errors are relatively small for the resolved scales, but grow in size for the unresolved scales. The intermediate ensemble (green dots) shows smaller errors at smaller scales than the small ensemble (blue) by a km or two.

7 Sensitivity of forecast error (co)variances to ensemble size

Bayesian data assimilation techniques such as those based on variational or ensemble methods make use of short-range forecast uncertainty to regularise the problem of estimating the state of a system from a set of observations. A reliable estimate of the forecast error covariances are then essential to provide near optimal initial conditions for NWP. Such an estimate is often obtained by means of a forecast ensemble, which is run either in off-line mode (for climatological estimates of error covariances),

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

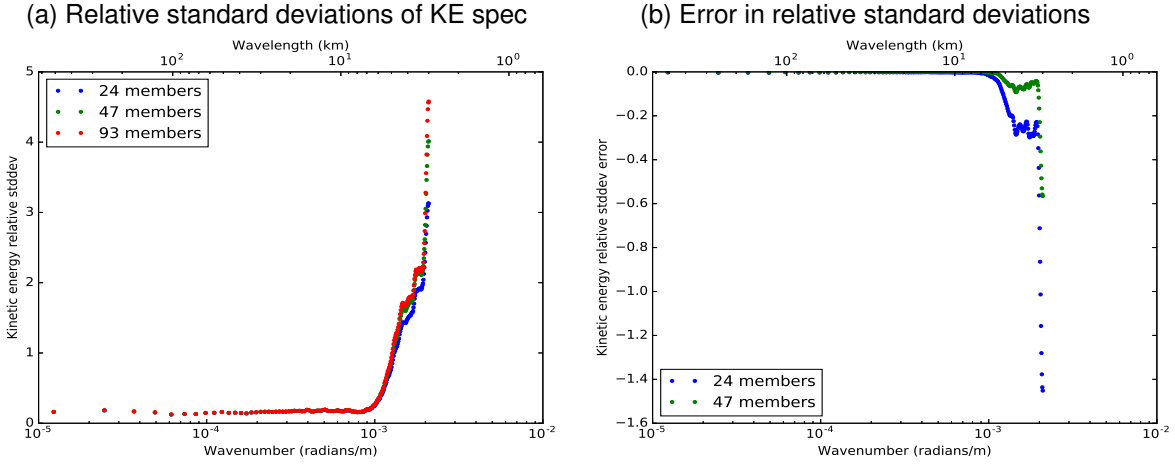

Figure 10. (a) The ratio $\bar{r}_E^{(N)}(k) = \bar{\sigma}_E^{(N)}(k)/\bar{E}^{(N)}(k)$ (the relative standard deviation of kinetic energy) for the small (blue), intermediate (green), and large (red) ensembles. (b) The error in $\bar{r}_E^{(N)}(k)$ for the small and intermediate ensembles. Computation of this error assumes that the large ensemble represents the true value, making the definition of error $\Delta\bar{r}_E^{(24/47)}(k) = \bar{r}_E^{(24/47)}(k) - \bar{r}_E^{(93)}(k)$. The small and intermediate ensemble results are based on 1000 sub-samples.

or in on-line mode as part of the forecast step of an ensemble-based data assimilation system (e.g. Buehner (2005); Pereira and Berre (2006); Schwartz and Liu (2014)). It is therefore important to understand how sampling error affects (co)variance estimates for different sized ensembles.

7.1 Estimates of forecast error variances and their sampling errors

The top panels of Figs. 11 and 12 are the sample variances of temperature (model level 36) and specific humidity (level 10) respectively, and in each case panel (a) uses the large ensemble and (b) uses the small ensemble. The temperature (T) variances in Fig. 11 are higher towards the NW corner of the area (well behind the cold front), than elsewhere, and the contrast is stronger for the large ensemble. Furthermore the large ensemble has variances of higher values and is smoother. The specific humidity (q) variances in Fig. 12 show large values close to the Peninsula in the SW part of the domain (at and just behind the cold front), and again the large ensemble has variances of higher values.

By assuming that the variances from the large ensemble contain negligible sampling error, we can study the sampling error in the variances computed from the smaller ensembles. We assume that a field of estimated forecast error variances sampled with an N-member ensemble, $\mathbf{v}^{(N)}$, is related to $\mathbf{v}^{(\infty)} \approx \mathbf{v}^{(93)}$, via

$$\mathbf{v}^{(N)} = \mathbf{v}^{(\infty)} + \mathbf{g}^{(N)}, \tag{2}$$

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

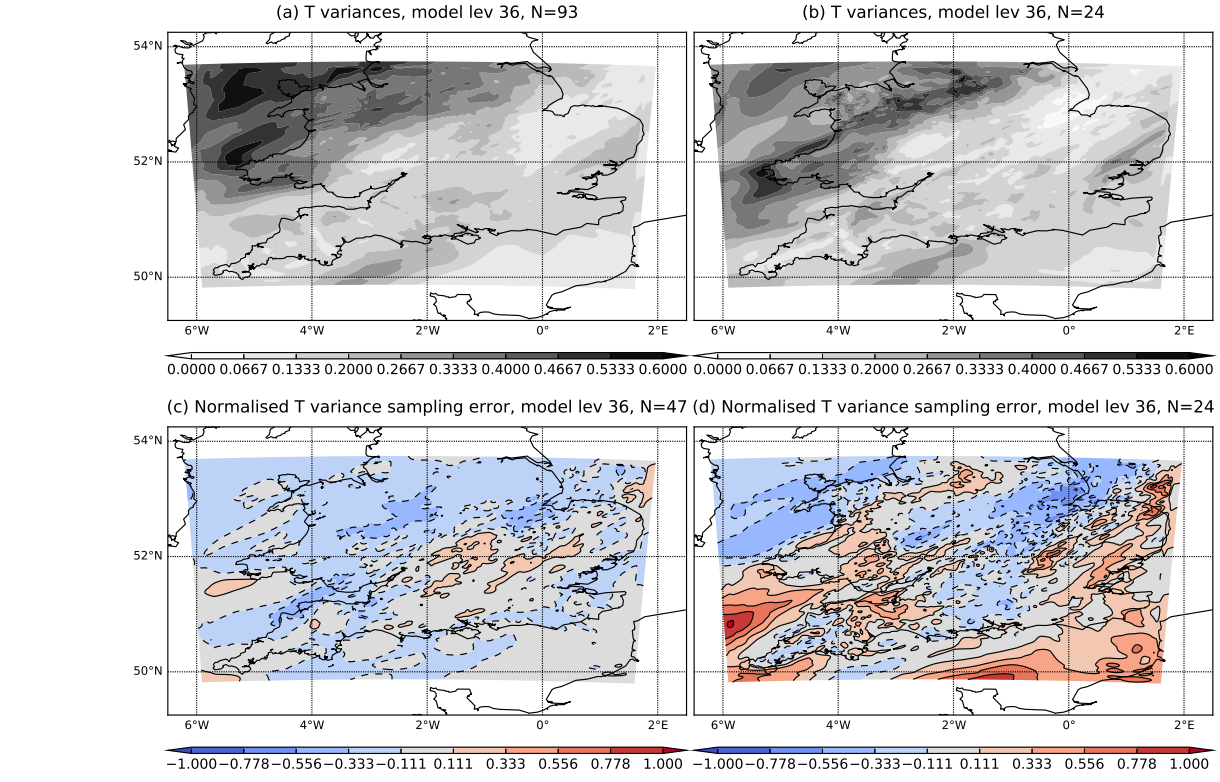

Figure 11. Top row: sample temperature variances (units K^2) computed from (a) the large, and (b) the small ensembles. Bottom row: estimates of the normalised and scaled sampling errors ($\mathbf{g}^{(N)\prime}$ as defined in (3)) for temperature variances for (c) the intermediate, and (d) the small ensemble. Plots are for model level 36 (~ 4.5 km above sea level), and for 20th September 2011 15Z. The intermediate and small ensembles each represents a single sample from the large ensemble. Negative values have dashed contours, and positive (negative) values are red (blue).

where $\mathbf{g}^{(N)}$ is the vector of sampling errors in an estimate of variance from an ensemble of N members[6]. Fields $\mathbf{v}^{(N)}$, $\mathbf{v}^{(\infty)}$, and $\mathbf{g}^{(N)}$ comprise the diagonal elements of the sample forecast error covariance matrix, $\mathbf{B}^{(N)}$, the sampling-error-free matrix, $\mathbf{B}^{(\infty)}$, and the sampling error matrix, $\mathbf{G}^{(N)}$, respectively. Let primed vectors correspond to the above, but normalised by the respective elements of $\mathbf{v}^{(\infty)}$:

$$\mathbf{v}^{(N)\prime} = \mathbf{1} + \mathbf{g}^{(N)\prime}, \tag{3}$$

where $\mathbf{1}$ is a vector of 1s[7]. The bottom panels of Figs. 11 and 12 are the estimates of $\mathbf{g}^{(N)\prime}$ for T and q respectively, and in each case panel (c) uses the intermediate ensemble and (d) uses the small ensemble. Positive (negative) values suggest that

[6]On notation: given a vector representation of a field, e.g. $\mathbf{g}$, the value at position r_i is denoted $g(r_i)$; and in spectral space, e.g. $\bar{\mathbf{g}}$, the value at a particular wavenumber is $\bar{g}(k_j)$. We swap between these notations depending upon which is the most appropriate in a given context.

[7]Note that $\mathbf{g}^{(N)\prime}$ is defined as the ratio between the sample variance error and the 'true' – rather than the sample – variance. This complies with the prescript "one should never, ever, put a random number in a denominator." (Penland, 2011).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

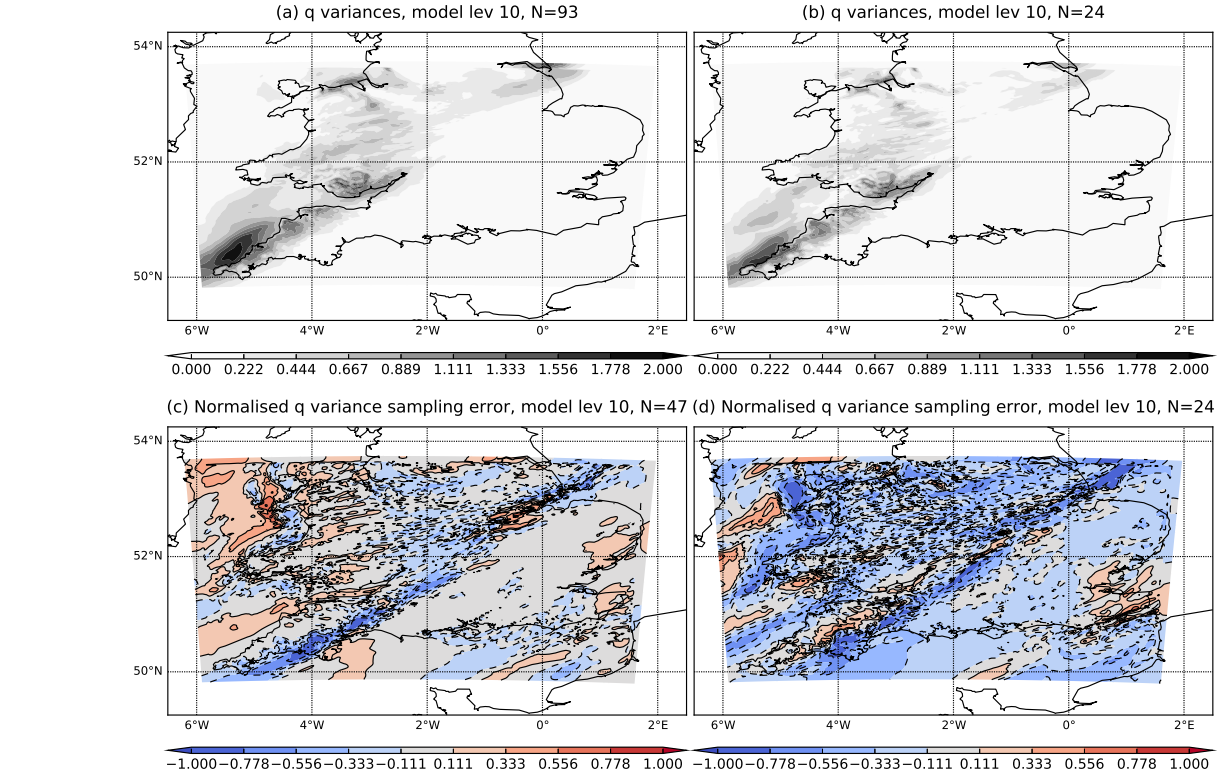

Figure 12. As Fig. 11 but for specific humidity (variance units $(\mathrm{g/kg})^2$) at level 10 ($\sim$ 500 m above sea level).

the sample variance is over-(under-)estimated and values close to ± 1 represent sampling errors that are of the same size as the 'true' variance. The advantage of showing $\mathbf{g}^{(N)\prime}$, instead of $\mathbf{g}^{(N)}$ is that the former shows up errors even when the actual variances are small.

Relative sampling errors for T are generally smaller in magnitude for the intermediate ensemble (Fig. 11 panel c) than for the small ensemble (panel d). The (probably) anomalously small T variances measured by the small ensemble (e.g. in the Irish Sea as seen in panel b) do show as negative values in Fig. 11d, although there are areas that have a larger magnitude of relative variance discrepancy. The story is similar for the the q variances measured by the small ensemble, with the relative sampling errors smaller in magnitude for the intermediate ensemble (Fig. 12 panel c) than for the small ensemble (panel d). These results (especially the different patterns produced by the different sized ensembles in Fig. 11) suggest that increasing the number of ensemble members in the way described in Sect. 3.2 does create members with a degree of independence (i.e. the new members are not linearly related to the existing members), thus supporting the result of Fig. 3.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

7.2 The covariances of sampling errors in the variances

The covariance of the above normalised variance errors can be written in the following way by adapting a result given as Eq. (6) in Raynaud et al. (2009) (assuming that the sample variance estimates are unbiased):

$$\mathrm{cov}\left[\mathbf{g}^{(N)\prime}\right] = \frac{2}{N-1}\mathbf{C}^{(\infty)} \circ \mathbf{C}^{(\infty)} \equiv \frac{2}{N-1}\mathbf{C}_{\mathbf{g}}, \tag{4}$$

where $\mathbf{C}^{(\infty)}$ is the noise-free forecast error correlation matrix, and $\circ$ denotes the Schur product. In (4), the correlation matrix of variance sampling errors, $\mathbf{C}_{\mathbf{g}}$, has been defined as $\mathbf{C}_{\mathbf{g}} \equiv \mathbf{C}^{(\infty)} \circ \mathbf{C}^{(\infty)}$ (note that the Schur product of two correlation matrices is also a correlation matrix (Gaspari and Cohn, 1999)). This result confirms that sampling error in the variance is expected to reduce with increasing N, and will have shorter length-scales than for forecast error correlations (due to the Schur product).

By defining another normalisation $\mathbf{g}^{(N)\prime\prime} = \left(\sqrt{(N-1)/2}\right)\mathbf{g}^{(N)\prime}$, Eq. (4) then informs us that the covariance $\mathrm{cov}\left[\mathbf{g}^{(N)\prime\prime}\right] = \mathbf{C}_{\mathbf{g}}$, whose elements can be estimated using the Wiener-Khinchin theorem. The Wiener-Khinchin theorem says that the correlation of $g^{(N)\prime\prime}(x_i)$ with $g^{(N)\prime\prime}(x_i + r_i)$ (call this auto-correlation $c_{\mathbf{g}}^{(N)}(r_i)$, which is the matrix element $(\mathbf{C}_{\mathbf{g}})_{x_i,x_i+r_i}$, here assumed to be independent of x_i for homogeneity) is:

$$\begin{aligned} c_{\mathbf{g}}^{(N)}(r_i) &= \frac{1}{n_x}\sum_{i'} g^{(N)\prime\prime}(x_{i'})g^{(N)\prime\prime}(x_{i'}+r_i) = \\ &= \frac{1}{n_x}\underbrace{\left(\frac{1}{n_x}\sum_j \left|\bar{g}^{(N)\prime\prime}(k_j)\right|^2 \exp\left(\frac{2\imath\pi ij}{n_x}\right)\right)}_{\text{inverse Fourier transform}}, \end{aligned} \tag{5}$$

where i, i' represent position index, j represents wavenumber index, $\imath = \sqrt{-1}$, and n_x is the number of points in the longitudinal direction. This theorem says that $c_{\mathbf{g}}^{(N)}(r_i)$ is proportional to the inverse Fourier transform of the power spectrum $S^{(N)\prime\prime}(k_j) \equiv \left|\bar{g}^{(N)\prime\prime}(k_j)\right|^2$, where $\bar{g}^{(N)\prime\prime}(k_j)$ is the Fourier transform of $g^{(N)\prime\prime}(x_i)$.

The top panels of Fig. 13 show the power spectra $S^{(N)\prime\prime}(k_i)$ for T (panel a) and q (b) for the small (blue) and intermediate (green) ensembles. Similar to the processing done for Sect. 6, the spectra are averaged latitudinally, and then vertically (for T this is between vertical levels ~ 3 and ~ 9 km above sea level, and for q this is below ~ 3 km), but they are valid for the same time as the plots in Figs. 11 and 12 (15Z). Before the Fourier transform, the fields are de-trended in the way described in point 2 of the first numbered list of Sect. 6. Since $\mathbf{C}_{\mathbf{g}}$ is in principle a constant matrix (i.e. independent of N) the spectra should in principle be the same – any significant deviation being due to departures from the assumptions made for (4). The intermediate ensemble has systematically lower power than the small ensemble. This questions whether the approximation $\mathbf{v}^{(\infty)} \approx \mathbf{v}^{(93)}$ used for (3) is well justified, (although we still assume that it is useful).

In order to assess any decrease in sampling noise when moving from the small to the intermediate ensemble, the middle panels of Fig. 13 plot the ratio $S^{(47)\prime}(k_j)/S^{(24)\prime}(k_j)$, where $S^{(N)\prime}(k_j) \equiv \left|\bar{g}^{(N)\prime}(k_j)\right|^2$ is the spectral weight of the variance sampling errors without the $\sqrt{(N-1)/2}$ normalisation. These show a similar picture for T (panel c) and q (d), where increasing the number of ensemble members from (23+1) to (46+1) reduces sampling error in all scales (including unresolved scales) to around a third in power. This decrease in sampling error is greater for the larger scales than for the smaller scales.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

The bottom panels of Fig. 13 show the correlation functions, $c_{\mathbf{g}}^{(N)}(r_i)$ from (5), found using the procedure using the inverse Fourier transform mentioned above for the small (blue lines) and intermediate (green lines) ensembles. These functions have a similar shape for T (panel e) and q (f), although q variance errors for the small ensemble do show a slight fluctuation in the correlation at around 200 km separation. The length-scale for variance errors in T is longer than that for q. If all assumptions made above are satisfied – that variance errors are additive as in (2), that $\mathbf{v}^{(\infty)} \approx \mathbf{v}^{(93)}$, and that (4) holds – a calculation of the correlation would satisfy $c_{\mathbf{g}}^{(N)}(r=0) \approx 1$ (the fact that this is not an exact equality is due to sampling error). Applying (5), the value $c_{\mathbf{g}}^{(N)}(r_i=0)$ is as follows:

$$c_{\mathbf{g}}^{(N)}(0) = \frac{1}{n_x}\sum_i \left(g^{(N)\prime\prime}(x_i)\right)^2 \approx \sigma^{\mathbf{g}^{(N)\prime\prime}2} + \left\langle g^{(N)\prime\prime}\right\rangle^2, \tag{6}$$

where $\sigma^{\mathbf{g}^{(N)\prime\prime}}$ and $\left\langle g^{(N)\prime\prime}\right\rangle$ are the sample standard deviation and mean of $g^{(N)\prime\prime}(x_i)$ respectively. In the limit $N \to \infty$, we expect $\left\langle g^{(N)\prime\prime}\right\rangle^2 \to 0$, and (4) tells us that $\sigma^{\mathbf{g}^{(N)\prime\prime}} \to 1$, leading to $c_{\mathbf{g}}^{(N)}(0) \to 1$. The observation that $c_{\mathbf{g}}^{(N)}(0)$ does not approach unity in Fig. 13e and f as N increases indicates that one or more of the assumptions mentioned above is not true, including that $\mathbf{v}^{(\infty)} \approx \mathbf{v}^{(93)}$. We regard this diagnostic as a new potential test of whether an ensemble (in this case the large ensemble) is large enough or not. We find that for T the values are $c_{\mathbf{g}}^{(N)}(0) < 1$ for both of the smaller ensembles. Interestingly, this may suggest that $\mathbf{v}^{(93)} > \mathbf{v}^{(\infty)}$ (connected with the step going from (2) to (3)).

7.3 Exponential fit to the correlation functions

Various mathematical forms of correlation function are fitted to the above empirical functions for $c_{\mathbf{g}}^{(N)}(r_i)$: a Gaussian ($c_{\mathrm{Gauss}}(r_i) = a\exp\left(-(r_i-\mu)^2/(2L^2)\right)+b$), an exponential ($c_{\exp}(r_i) = a\exp\left(-|r_i-\mu|/L\right)+b$), and a second-order auto-correlation function $c_{\mathrm{SOAR}}(r_i) = a(1+|r_i-\mu|/L)\exp\left(-|r_i-\mu|/L\right)+b$, where a, μ, L, and b are positive constants (see Eq. (2.34) of Gaspari and Cohn (1999)). Of these three forms, the exponential provided the best fit (by eye the fits are very close with the exception of the fluctuations of the blue curve in Fig. 13f for lengths greater than ~ 100 km).

Note that the exponential form is useful because it allows for such errors in the variance to be modelled with a first order autoregressive process (see e.g. Sect. 11.3 of Evensen (2009)). Applied to $g^{(N)}(x_i)$ (i.e. unprimed as in (2)), errors at subsequent points in the longitudinal direction can be modelled by

$$g^{(N)}(x_i) = \varrho g^{(N)}(x_{i-1}) + \sqrt{\frac{2(1-\varrho^2)}{N-1}}v^{(\infty)}(x_i)\epsilon(x_i), \tag{7}$$

where ϱ is the exponential correlation $\varrho = \exp(-\Delta x/L)$, $\Delta x = 1.5$ km, $v^{(\infty)}(x_i)$ is the error-free variance, and $\epsilon(x_i)$ is white noise with zero mean and unit variance. Equation (7) may be useful in other work to simulate variance fields for an N-sized ensemble (for a given 'true' variance field).

The best-fit parameters a, μ, L, and b of the exponential change according to the number of ensemble members used to compute the variance error. Figure 14 shows how a (top row) and L (bottom row) change with the number of ensemble members for T (left) and q (right) for $5 \le N \le 47$. In Fig. 14 the mean and standard deviation (error bars) of the best fit have been calculated over the 1000 sub-samples of each N-member ensemble. If the intermediate ensemble $N = 47$ is large enough,

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

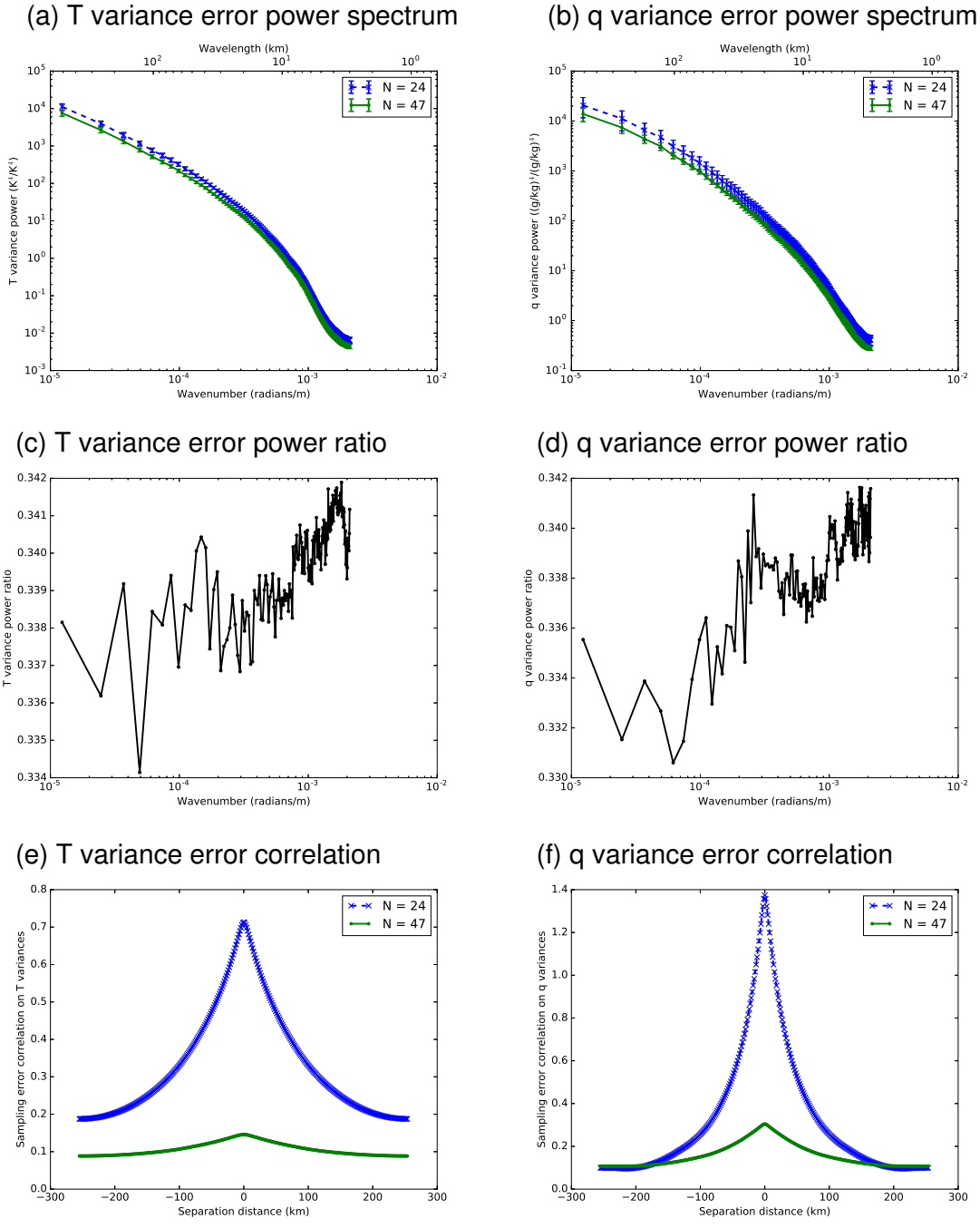

Figure 13. Top row: power spectra for estimates of sampling variance errors, $S^{(N)''}(k_j)$ for temperature (left-hand panels) and specific humidity (right-hand panels). Calculations are made for the small (blue), and the intermediate (green) ensembles. Middle row: ratios of $S^{(47)'}(k_j)/S^{(24)'}(k_j)$. Bottom row: correlation functions $c_{\mathbf{g}}^{(N)}(r_j)$. Calculations are valid for 20th September 2011 15Z and the ensembles are sub-sampled 1000 times from the large ensemble.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

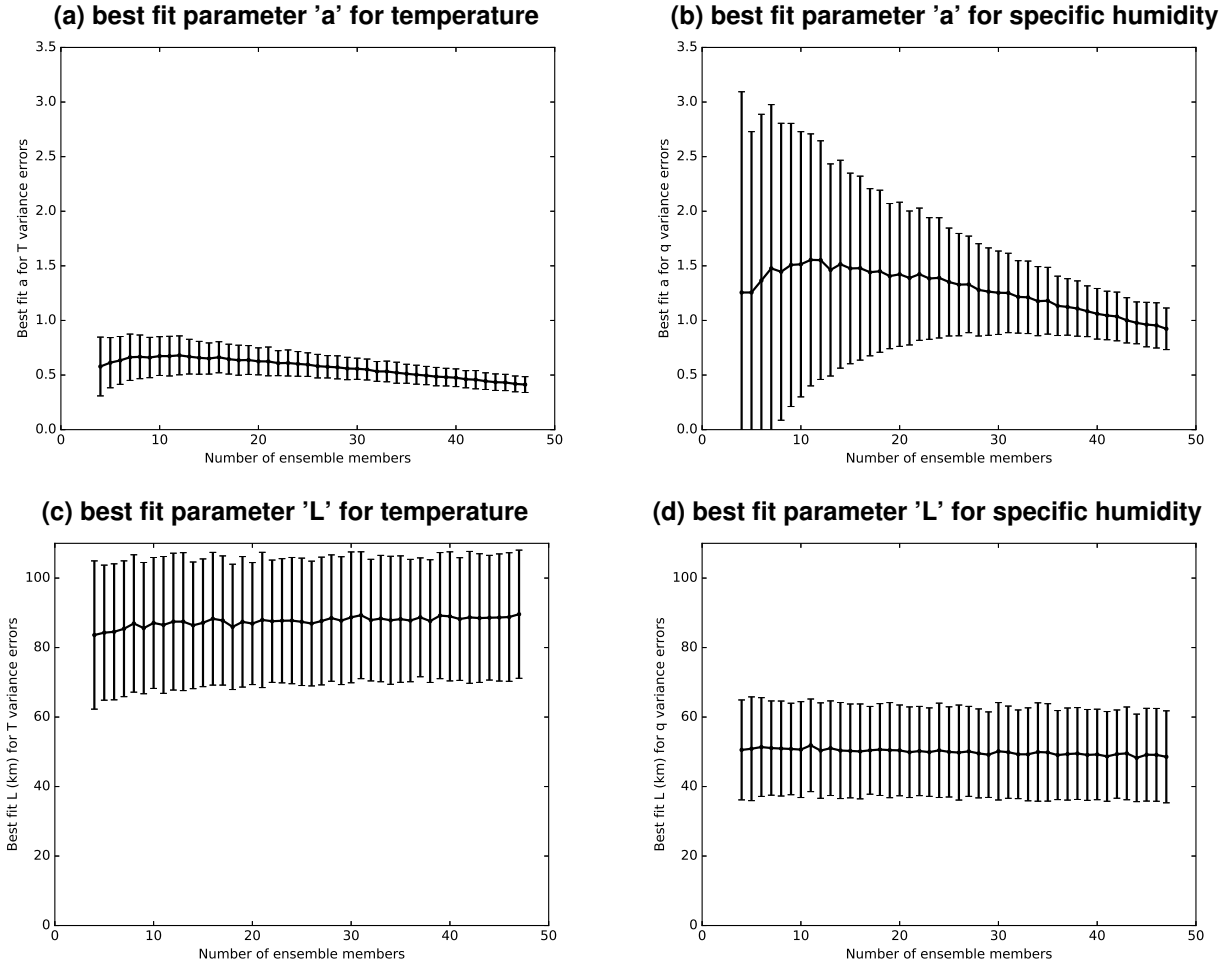

Figure 14. Best fit parameters a (top row) and L (bottom row) for the exponential form of the correlations of variance sampling error $c_{\exp}(r_i) = a\exp(-|r_i - \mu|/L) + b$ as a function of the number of ensemble members, N. The left panels are for temperature and the right panels are for specific humidity. Note that each row has the same y-axis scale to allow easy comparison of parameter values between temperature and specific humidity. Values are computed for each N using 1000 sub-samples of N members from the 93. The parameters μ and b are found to be small (not shown) and calculations are valid for 20th September 2011 15Z.

we would expect parameters to converge to a value with a small standard deviation. Convergence appears to have been reached for L for both T and q, although the standard deviations do not reduce significantly. Convergence appears not to have been reached for a for both T and q, although the standard deviations do reduce significantly. Based on these results for these test data it is difficult to judge whether the intermediate ensemble is 'large enough' to neglect sampling error in this context. An ensemble even larger than 93-members (as a proxy of the 'true' statistics) would be required to make a similar assessment up to 93 members.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

7.4 Estimates of forecast error correlations and their sampling errors

The above arguments may be extended to the analysis of sampling errors in correlations. The analogue of (2) for covariances is as follows:

$$\mathbf{B}^{(N)}_{\mathbf{x}\mathbf{x}'} = \mathbf{B}^{(\infty)}_{\mathbf{x}\mathbf{x}'} + \mathbf{G}^{(N)}_{\mathbf{x}\mathbf{x}'}, \tag{8}$$

where the matrix element indices $\mathbf{x}$ and $\mathbf{x}'$ correspond to positions. Dividing each side by $\sqrt{v^{(\infty)}_{\mathbf{x}} v^{(\infty)}_{\mathbf{x}'}}$, and letting primed versions of the matrix elements be such normalised versions of symbols in (8), gives the analogue of (3):

$$\mathbf{B}^{(N)\prime}_{\mathbf{x}\mathbf{x}'} = \mathbf{B}^{(\infty)\prime}_{\mathbf{x}\mathbf{x}'} + \mathbf{G}^{(N)\prime}_{\mathbf{x}\mathbf{x}'}, \tag{9}$$

where $\mathbf{B}^{(\infty)\prime}_{\mathbf{x}\mathbf{x}'}$ is the error-free background error correlation[8]. The variance of the normalised covariance error, $\mathbf{G}^{(N)\prime}_{\mathbf{x}\mathbf{x}'}$, can be written in the following way by adapting a result given as Eq. (10) in Ménétrier et al. (2014):

$$E\left[\mathbf{G}^{(N)\prime 2}_{\mathbf{x}\mathbf{x}'}\right] = \frac{1}{N-1}\left(1 + \mathbf{B}^{(\infty)\prime 2}_{\mathbf{x}\mathbf{x}'}\right). \tag{10}$$

The top row of Fig 15 are maps of ensemble mean T at model level 36 (~ 4.5 km above sea level) and q at model level 11 (~ 500 m above sea level). For T (panel a) the air ahead of the cold front (SE corner) is about 8K warmer that the air behind the front at this level, and for q (panel b) a strip of moist air divides regions ahead of the front (moister by ~ 2 g/kg than air ahead of the front at this level) from air behind the front (moister by ~ 5 g/kg than air behind the front). The front is more advanced (further SE) at higher levels than at lower levels, which is characteristic of the structure of a mid-latitude cold front.

The second row are maps of the auto-correlations of forecast errors of T and q denoted as $\mathbf{B}^{(93)\prime}_{\mathbf{x}\mathbf{x}'}$ in (9) (see also footnote 8). The auto-correlations are between the position of the cross and the rest of the domain. These correlations are computed from the large ensemble, and we assume that they are close to the error-free correlations. The error correlation patterns are aligned with the front. The clear line of zero correlation in panel (d) probably marks the front's exact position near the surface. The correlation length-scales for T errors are generally larger than those for q errors, which is consistent with the findings for variance error correlation length-scales for these quantities shown in Fig. 13 (panels e and f) and in Fig. 14 (panels c and d), given the relationship between the statistics of forecast and variance errors in (4).

The third and fourth rows are maps of $\mathbf{B}^{(47)\prime}_{\mathbf{x}\mathbf{x}'}$ and $\mathbf{B}^{(24)\prime}_{\mathbf{x}\mathbf{x}'}$ respectively. Note that there is no guarantee that $\mathbf{B}^{(N<93)\prime}_{\mathbf{x}'\mathbf{x}'} = 1$ since the normalising variances are here always computed from the large ensemble – see (9). These maps show that the large-scale structure of the second row is maintained, but the progressively larger magnitude of sampling error as N is decreased – as in (10) – introduces small-scale noise and more irregular demarcation lines.

7.5 Estimation of forecast error correlation length-scales

The final set of diagnostics to examine are the correlation length-scales, derived from the large, intermediate, and small ensembles. These length-scales are found in the following way:

[8]It may be natural to use the notation $\mathbf{C}^{(N)}_{\mathbf{x}\mathbf{x}'}$ to represent the correlation denoted as $\mathbf{B}^{(N)\prime}_{\mathbf{x}\mathbf{x}'}$ in (9), but we continue with the primed notation to mean 'normalised with the error-free standard deviations' for consistency with Sect. 7.1.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions
Open Access
EGU

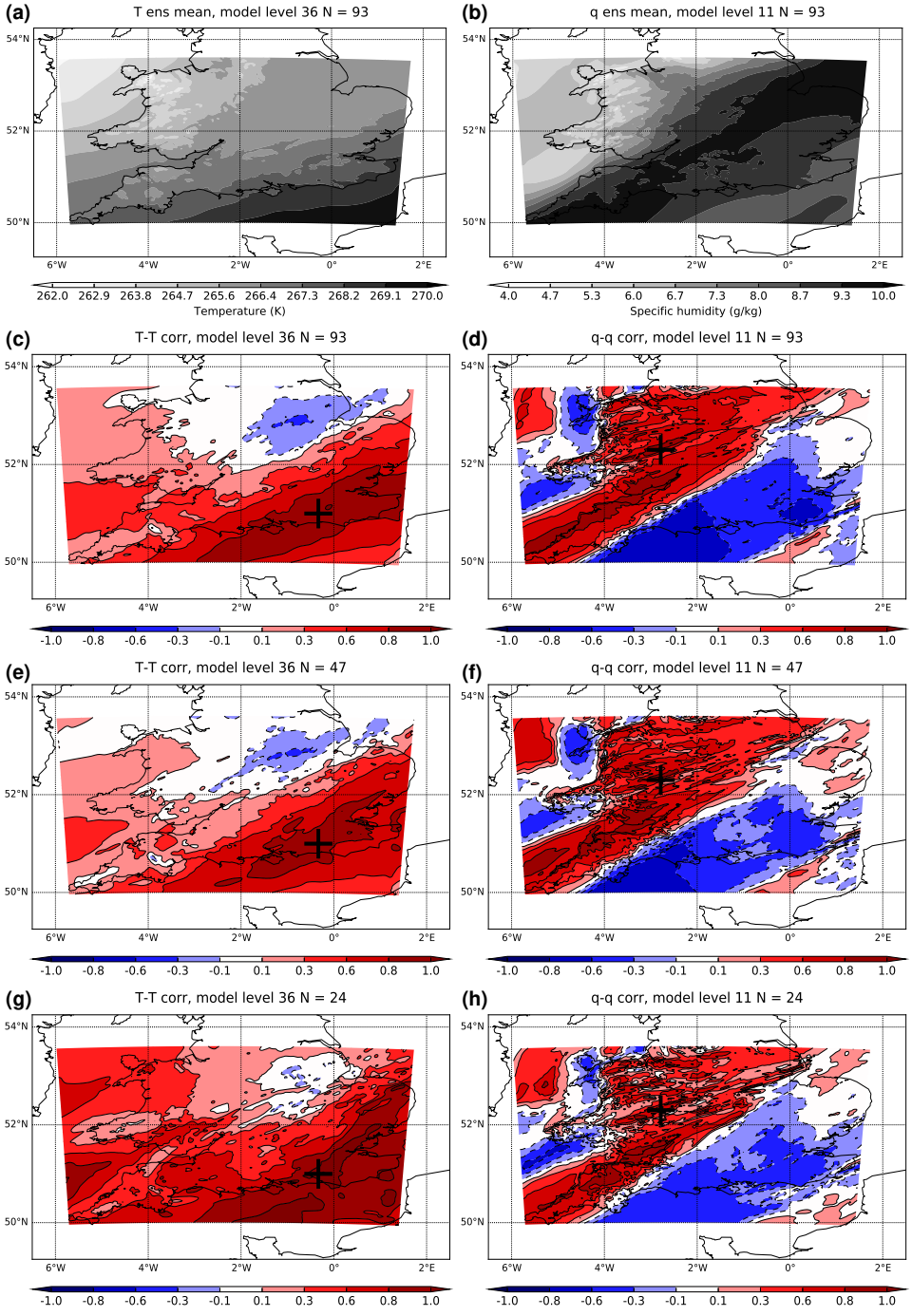

Figure 15. Top row: ensemble mean temperature (left) and specific humidity (right) computed from the large ensemble. Second row: spatial patterns of the sample error correlation, $\mathbf{B}_{\mathbf{xx}'}^{(93)'}$ between temperature error at the cross, $\mathbf{x}'$, with temperature error elsewhere, $\mathbf{x}$ (left), and for specific humidity (right) computed from the large ensemble. Third (fourth) row: similar sample correlations to the second row, but computed from the intermediate (small) ensemble, $\mathbf{B}_{\mathbf{xx}'}^{(47)'}$ ($\mathbf{B}_{\mathbf{xx}'}^{(24)'}$), and where the standard deviations are from the large ensemble (proxy truth). The crosses are at 0.33 E, 51 N, level 47 (left panels) and 2.79 E, 52.3 N (right panels). Calculations are valid for 20th September 2011 15Z. The intermediate and small ensembles each represents a single sample from the large ensemble. Positive (negative) values are red (blue).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

1. For each point in the longitude/latitude domain, $\mathbf{x}'$, compute the sample auto-covariance between the forecast error of T or q at $\mathbf{x}'$ and the same quantity at surrounding points, $\mathbf{x}$ within a local region of 30 grid boxes in the longitude and latitude directions either side of $\mathbf{x}'$. For a given fixed $\mathbf{x}'$ this gives the 2D covariance field $\mathbf{B}^{(N)}_{\mathbf{x}\mathbf{x}'}$ (a function of $\mathbf{x}$).

2. Estimate the correlation by dividing by the field $\sqrt{v^{(\infty)}_{\mathbf{x}} v^{(\infty)}_{\mathbf{x}'}}$ in the square region again using the approximation $v^{(\infty)}_{\mathbf{x}} \approx v^{(93)}_{\mathbf{x}}$. This results in a field that is about unity at $\mathbf{x}'$, and mostly decays with distance from $\mathbf{x}'$.

3. Fit the correlation field to a two-dimensional exponential of the form:

$$a \exp\left(-\left|\frac{(x-x')\cos\theta + (y-y')\sin\theta}{L_1}\right|\right) \times \exp\left(-\left|\frac{(y-y')\cos\theta - (x-x')\sin\theta}{L_2}\right|\right) + b, \quad (11)$$

by varying the parameters x', y', a, b, θ, L_1, and L_2. Here $(x,y) = \mathbf{x}$, a is the amplitude of the exponential function, b is its offset, θ is its orientation (from lines of constant latitude), and L_1 and L_2 are the length-scales along the principle axes. An exponential form may not be the ideal choice describing the shape of a correlation function, but we assume that the good fit of exponential functions to the variance error correlations in Sect. 7.3 carries through to forecast error correlations.

Our interest is in the quantity $\max(L_1, L_2)$. Maps of this quantity as a function of $\mathbf{x}'$ are made for each ensemble size considered. Even though for the smaller ensembles this diagnostic would benefit from averaging over 1000 sub-samples, only one sub-sample is used as the computations are very time consuming.

Figure 16 shows fields of $\max(L_1, L_2)$ for T (level 36, left panels), and q (level 11, right) for the large (panels a and b), the intermediate (c and d), and the small (e and f) ensembles. All panels have similar large-scale patterns that have alignment with the front. Temperature length-scales for the large ensemble (a) are generally shorter along the strip from the NE of the domain down to the SW, than at other regions, with the smallest values over the East Midlands (~ 80 km half-width). The largest values reach ~ 140 km. Note that on this model level and at this time most of the air is probably behind the front. Specific humidity length-scales (b) have a more complicated structure with the smallest values over parts of Wales (~ 20 km half-width, probably associated with the varying orography), and the largest values reach ~ 150 km over the Peninsula in the SW part of the domain. Given the results of Sects. 7.2 and 7.3 it is perhaps surprising that the maximum length-scales of q are found to be longer than those of T. These results need not be inconsistent though as the previous results indicate the average length-scales. The average length-scale for q is brought down by contributions from the very small length-scales of q in in the NW parts of panel (b).

As the number of members is reduced to (46+1) (panels c and d), and then to (23+1) (e and f), the position-dependent length-scales change. For T, reducing the ensemble size progressively underestimates the length-scales where the 'true' length-scales are short, and overestimates them when the 'true' length-scales are long. Not only does this show that error covariance length-scale can be significantly affected by sampling noise, it is an interesting result in itself. As for previous results, without having an even larger ensemble, it is difficult to judge for sure whether the large ensemble is 'large enough' to be represent the true length-scales accurately.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

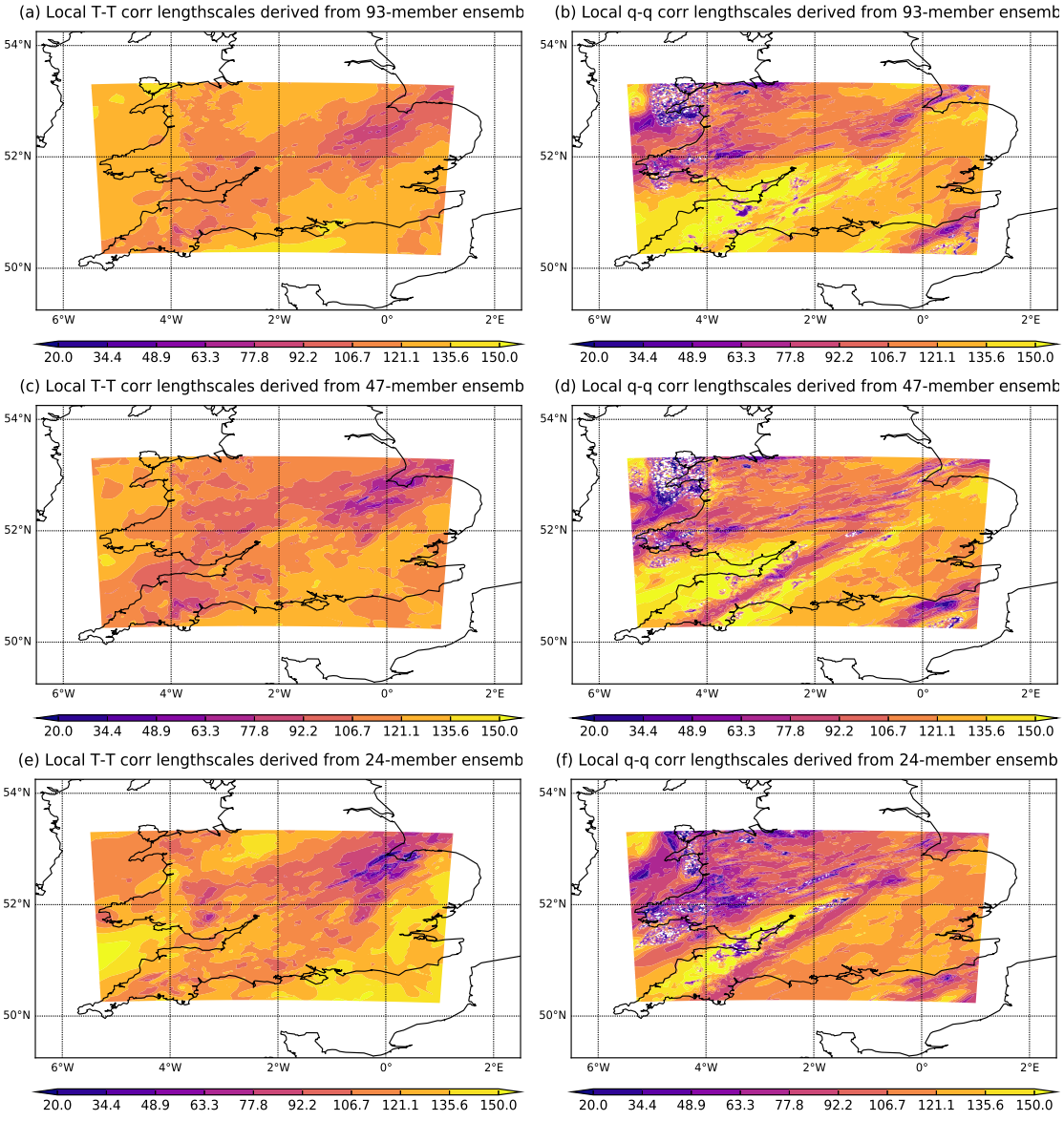

Figure 16. Local temperature (model level 36, left panels) and specific humidity (level 11, right) correlation length-scales ($\max(L_1, L_2)$, where L_1 and L_2 are defined in (11)). Calculations are valid for 20th September 2011 15Z. The intermediate and small ensembles each represents a single sample from the large ensemble.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

Geoscientific Model Development Discussions

8 Conclusions

The aim of this work is to first generate a 'large' convective-scale ensemble and then to develop a suite of diagnostics to help understand how a large ensemble adds degrees of freedom to an existing smaller ensemble prediction system. Many of the diagnostics have, to the authors' knowledge, not been applied before to study sampling error. An ensemble of $N = (92+1)$ members (nearly four times larger than the (23+1) operational ensemble members at the time that the integrations were made[9]) was generated for use in a convective-scale limited area model of the Southern UK (SUK). These diagnostics are applied to the test ensemble for illustration. The case study chosen has rain band structures that are of interest to the DIAMET project. The ultimate aim of the diagnostics when applied to a more substantial set of data is to help decide whether the number of ensemble members is 'enough' to neglect sampling error, or at least to help practitioners get a feel for how much sampling error is present. The authors have the application of high-resolution models and precipitation prediction systems in mind, although most of the diagnostics are useful to a wider range of models.

For the purposes of this study it is assumed that the large ensemble gives reference diagnostics that are free of sampling errors, which presents an opportunity to estimate sampling errors of smaller ensemble sizes. Smaller ensemble sizes are studied by choosing members at random from the large ensemble, and the diagnostics then averaged over a set of such sub-samples. Predominantly used were a (23+1)-member ensemble (the 'small' ensemble), and a (46+1)-member ensemble (the 'intermediate' ensemble). Some diagnostics use only a single random sub-sample, as averaging over a large number of sub-samples would give diagnostics similar to those of the large ensemble. For other diagnostics, it was meaningful and practical to average over a larger number of sub-samples (1000). Only one case study is used in this paper, which is due to limited resources. We acknowledge that a single case is insufficient to make definitive conclusions regarding any system, but here it does serve to illustrate the methodology and the diagnostics developed. We now re-visit the aims as they are set out towards the end of Sect. 1.

How can linearly independent extra members be generated from an existing ensemble?

The large ensemble is developed from an existing small ensemble by creating additional negative versions of the existing perturbations in a set of nested models (23+1)→(46+1)→(92+1), and spinning-up the resulting model states (Sect. 3 and Fig. 2). The negative perturbations of course are not immediately linearly independent of the existing perturbations, but our expectation that a degree of independence develops as the ensemble of non-linear model runs evolve is confirmed (Fig. 3).

How does the ensemble size impact the probabilistic forecasts of rainfall?

The purpose of studying how N changes the probabilistic forecast of rainfall is to gain a sense of how much information is added with the extra members. It is useful to do this for a range of precipitation rate thresholds to test an ensemble's ability to predict a sufficient range of precipitation outcomes that match the observations. In our case study the large, intermediate, and

[9]The Met Office's operational convective-scale ensemble currently has (11+1) members, and is produced by downscaling members from the global ensemble (Hagelin et al., 2007).

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

small ensembles forecast non-zero probabilities of rain along rain bands 1 and 2 that were positioned inside the SUK domain (Fig. 1), but a zero probability of heavy rain in band 2, even though in reality the rain is heavy there (Figs. 1 and 8). The spread of rainfall outcomes affects the probability diagnostics, and our test ensembles are found to be over-spread and biased (Fig. 7), and increasing the ensemble size does not improve the bias of the rainfall seen in the small ensemble. Adding extra members does increase the variance of the (already too high) rainfall rates as expected (Fig. 5), but this does not always appear in other indicators of spread, such as the interquartile range (Fig. 6), which actually sometimes shows a decrease in the spread for the large ensemble. This latter finding is a non-Gaussian effect and serves to show that one should not rely on the variance alone as an indication of the spread.

The conclusions covering the above two questions suggest that adding extra members in this system does not have a positive impact on the diagnostics shown. On a positive note though, they do suggest that the method of generating the extra members produces results that are consistent with the behaviour of the original 24 members.

How does the large ensemble resolve the kinetic energy spectrum?

The KE spectrum contains information concerning the nature of the turbulence in the model – i.e. whether the bulk behaviour is characteristic of a 2D system with a k^{-3} kinetic energy (KE) spectrum (as seen in mid-latitude large-scale flows), or of a 3D turbulent system with a $k^{-5/3}$ KE spectrum (as might be expected at smaller scales). It is also able to provide information about the effective resolution of the system, the scales below which the spectrum deviates from its bulk character, and where the estimated uncertainty grows considerably. By looking at estimated relative errors in the KE spectrum provides further information on how much sampling error contributes to degrading the effective resolution.

The large ensemble studied appears to show behaviour more characteristic of a 2D system rather than of a 3D turbulent system (Fig. 9). This may be due to the case study (which is not particularly convective), and to the small size of the domain, not allowing enough opportunity for large-scale information from the boundaries to develop into full 3D motion. We estimate that the effective spatial resolution of the large ensemble is ~ 8 km (Fig. 9). Interestingly, this agrees roughly with Pielke's definition of model resolution as four or more grid lengths (Pielke, 2001). The sampling error in the relative spread of the KE spectrum (KE spread error divided by ensemble mean KE) is computed for the small and intermediate ensembles. This shows that increasing the ensemble from 24 to 47 members is associated with an increase of effective resolution of 1-2 km and a reduction of sampling error at all scales (Fig. 10).

How does ensemble size affect the estimates of (co-)variability of thermodynamic and moisture fields?

There are many diagnostics that are straightforward to produce and to interpret to help study the effect of sampling error in variance and covariance estimates. This includes the computation of sample variances, relative errors in variances, and spatial correlations as a function of N. These simple diagnostics are developed further. This includes the diagnosis of how aspects like correlation length-scales, and their uncertainties, change with N (including the domain-averaged and position-dependent length-scales). It also includes studying the sampling errors in a spectral fashion to help understand how increasing the ensemble size affects errors at difference scales. The shape of the spatial correlation functions can also be studied as a

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

function of N. Once the shape is determined, its fitting parameters (amplitude, length-scale, etc.) can be studied as a function of ensemble size to look for possible convergence to their 'true' values. Although these diagnostics are applied to thermodynamic and moisture fields they are not restricted to these kinds of fields.

For our case study there is a clear overall increase in the variances and their smoothness, and a reduction in the relative variance sampling error as N is increased (Figs. 11 and 12). Examining relative variance error, rather than just variance error, is useful because it highlights sampling errors even when the actual variances are small.

Studying the power spectra of T and q relative variance errors showed that adding the extra ensemble members affected these quantities in similar ways. All scales (resolved and unresolved) are improved when going from the small to the intermediate ensemble, but more so for the large scales (Figs. 13c and d). The correlation of sampling errors in variance as a function of distance follows an exponential form (Figs. 13e and f). The parameters that describe the exponential are estimated by doing a least-squares fit to the correlations computed from the ensembles. The two most relevant parameters (the amplitude and the length-scale of the correlation functions), change as a function of ensemble size, but it is not clear whether their values have converged over the ranges of N studied (Fig. 14). Even though these (domain averaged) length-scales do not change much with ensemble size over the ranges studied, the position-dependent length-scales do (Fig. 16). The correlation patterns in T and q do show reduced noise as the number of members is increased, which confirms a very well known result (Fig. 15).

Is the number of ensemble members in an ensemble enough to neglect sampling error?

It is difficult to say in this study whether it is reasonable to assume that sampling error could be neglected in the large ensemble. One particular diagnostic developed is Eq. (6), which should tend to unity as N increases provided certain conditions are met, including that the variance of the large ensemble has no sampling error. Applying this diagnostic to the data suggests that this is not the case (Figs. 13e and f), pointing towards the need to study this further with larger ensembles. That said there are visible reductions in sampling error in many of the fields.

Final comments

Although it is difficult to say for sure whether 93 members are enough to bring sampling errors to sufficiently small values, the fact that there are improvements means that the extra members do contain genuinely independent and useful information. The reduction in sampling error would benefit data assimilation applications, but the lack of sensitivity on other aspects, like rainfall probabilities, and biases suggests that the quality of probabilistic forecasts would not be improved. It is unwise to draw general conclusions though using only our particular (single case) test data.

There is still a need to study the statistics of ever larger ensembles (in convective- and large-scale EPSs). Various centres around the world have pushed the number of ensemble members to very large values (e.g. operationally N =256 at Environment Canada (Buehner et al., 2015; Caron et al., 2015), and non-operationally $N = 256$ (Yashiro et al., 2016) and even $N = 10240$ (Kondo and Miyoshi, 2016) at RIKEN in Japan). The models used in these systems do not approach convective-scales though, so there is still much progress to be made to allow large ensembles to be used routinely with convective-scale models. It is hoped that the new diagnostics presented in this paper will be adopted by others in future studies.

Code availability.

A selection of the software used in this paper is open-source and freely available on a Git Hub repository DOI 10.5281/zenodo.1013435

Author contributions. ACR and LHB adapted and ran the Met Office system to generate and integrate the extra members, and helped decide on some of the diagnostics used in the paper while they worked on the DIAMET project. SM and RNB developed and adapted the suite of diagnostics used in this study. SM provided an outline of the paper and RNB finalised the diagnostics and completed the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Disclaimer.

Acknowledgements. The authors would like to thank Neill Bowler, Jean-Francois Caron, Terry Davies, Jonathan Flowerdew, Stephen Pring, and Warren Tennant from the Met Office, and Roger Brugge and Peter Jan van Leeuwen from the University of Reading. We would also like to thank the Met Office/NERC for use of the Met Office/MONSooN supercomputing facilities. This work was funded by DIAMET (NERC grant NE/I005234/1) and by the National Centre for Earth Observation.

Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2017-260
Manuscript under review for journal Geosci. Model Dev.
Discussion started: 20 November 2017

References

Ancell, B. C.: Nonlinear characteristics of ensemble perturbation evolution and their application to forecasting high-impact events, Weather and Forecasting, 28, 1353–1365, 2013.

Baker, L., Rudd, A., Migliorini, S., and Bannister, R.: Representation of model error in a convective-scale ensemble prediction system, Nonlinear Processes in Geophysics, 21, 19–39, 2014.

Ballard, S. P., Li, Z., Simonin, D., and Caron, J.-F.: Performance of 4D-Var NWP-based nowcasting of precipitation at the Met Office for summer 2012, Quarterly Journal of the Royal Meteorological Society, 142, 472–487, 2016.

Bannister, R.: A review of operational methods of variational and ensemble-variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 607–633, https://doi.org/DOI:10.1002/qj.2982, 2017.

Bannister, R. N., Migliorini, S., and Dixon, M.: Ensemble prediction for nowcasting with a convection-permitting model–II: forecast error statistics, Tellus A, 63, 497–512, 2011.

Ben Bouallègue, Z. and Theis, S. E.: Spatial techniques applied to precipitation ensemble forecasts: from verification results to probabilistic products, Meteorological Applications, 21, 922–929, 2014.

Bick, T., Simmer, C., Trömel, S., Wapler, K., Hendricks Franssen, H.-J., Stephan, K., Blahak, U., Schraff, C., Reich, H., Zeng, Y., et al.: Assimilation of 3D radar reflectivities with an ensemble Kalman filter on the convective scale, Quarterly Journal of the Royal Meteorological Society, 142, 1490–1504, 2016.

Bonavita, M., Raynaud, L., and Isaksen, L.: Estimating background-error variances with the ECMWF Ensemble of Data Assimilations system: some effects of ensemble size and day-to-day variability, Quarterly Journal of the Royal Meteorological Society, 137, 423–434, 2011.

Bonavita, M., Isaksen, L., and Hólm, E.: On the use of EDA background error variances in the ECMWF 4D-Var, Quarterly Journal of the Royal Meteorological Society, 138, 1540–1559, 2012.

Bouallegue, Z. B., Theis, S. E., and Gebhardt, C.: Enhancing COSMO-DE ensemble forecasts by inexpensive techniques, Meteorologische Zeitschrift, 22, 49–59, 2013.

Bouttier, F., Raynaud, L., Nuissier, O., and Ménétrier, B.: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX, Quarterly Journal of the Royal Meteorological Society, 143, 390–403, 2016.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E.: The MOGREPS short-range ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 134, 703–722, 2008.

Buehner, M.: Ensemble-derived stationary and flow-dependent background-error covariances: Evaluation in a quasi-operational NWP setting, Quarterly Journal of the Royal Meteorological Society, 131, 1013–1043, 2005.

Buehner, M., McTaggart-Cowan, R., Beaulne, A., Charette, C., Garand, L., Heilliette, S., Lapalme, E., Laroche, S., Macpherson, S. R., Morneau, J., et al.: Implementation of deterministic weather forecasting systems based on ensemble–variational data assimilation at Environment Canada. Part I: The global system, Monthly Weather Review, 143, 2532–2559, 2015.

Buizza, R. and Palmer, T.: The singular-vector structure of the atmospheric global circulation, Journal of the Atmospheric Sciences, 52, 1434–1456, 1995.

Buizza, R. and Palmer, T. N.: Impact of ensemble size on ensemble prediction, Monthly Weather Review, 126, 2503–2518, 1998.

Buizza, R., Petroliagis, T., Palmer, T., Barkmeijer, J., Hamrud, M., Hollingsworth, A., Simmons, A., and Wedi, N.: Impact of model resolution and ensemble size on the performance of an ensemble prediction system, Quarterly Journal of the Royal Meteorological Society, 124, 1935–1960, 1998.

Caron, J.-F.: Mismatching perturbations at the lateral boundaries in limited-area ensemble forecasting: A case study, Monthly Weather Review, 141, 356–374, 2013.

Caron, J.-F., Milewski, T., Buehner, M., Fillion, L., Reszka, M., Macpherson, S., and St-James, J.: Implementation of deterministic weather forecasting systems based on ensemble–variational data assimilation at Environment Canada. Part II: The regional system, Monthly Weather Review, 143, 2560–2580, 2015.

Clark, A. J., Kain, J. S., Stensrud, D. J., Xue, M., Kong, F., Coniglio, M. C., Thomas, K. W., Wang, Y., Brewster, K., Gao, J., et al.: Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble, Monthly Weather Review, 139, 1410–1418, 2011.

Corazza, M., Kalnay, E., Patil, D., Yang, S.-C., Morss, R., Cai, M., Szunyogh, I., Hunt, B., and Yorke, J.: Use of the breeding technique to estimate the structure of the analysis "errors of the day", Nonlinear Processes in Geophysics, 10, 233–243, 2003.

Ehrendorfer, M.: A review of issues in ensemble-based Kalman filtering, Meteorologische Zeitschrift, 16, 795–818, 2007.

Errico, R. M.: Spectra computed from a limited area grid, Monthly Weather Review, 113, 1554–1562, 1985.

Evensen, G.: Data assimilation: the ensemble Kalman filter, Springer Science & Business Media, 2009.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, 1999.

Gebhardt, C., Theis, S., Paulat, M., and Bouallègue, Z. B.: Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries, Atmospheric Research, 100, 168–177, 2011.

Gilmour, I., Smith, L. A., and Buizza, R.: Linear regime duration: Is 24 hours a long time in synoptic weather forecasting?, Journal of the atmospheric sciences, 58, 3525–3539, 2001.

Golding, B., Ballard, S., Mylne, K., Roberts, N., Saulter, A., Wilson, C., Agnew, P., Davis, L., Trice, J., Jones, C., et al.: Forecasting capabilities for the London 2012 Olympics, Bulletin of the American Meteorological Society, 95, 883–896, 2014.

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W.: The Met Office convective-scale ensemble, MOGREPS-UK, Quarterly Journal of the Royal Meteorological Society, 2007.

Hamill, T. M.: Interpretation of rank histograms for verifying ensemble forecasts, Monthly Weather Review, 129, 550–560, 2001.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, Monthly Weather Review, 129, 2776–2790, 2001.

Harnisch, F. and Keil, C.: Initial conditions for convective-scale ensemble forecasting provided by ensemble data assimilation, Monthly Weather Review, 143, 1583–1600, 2015.

Hollan, M. A. and Ancell, B. C.: Ensemble Mean Storm-Scale Performance in the Presence of Nonlinearity, Monthly Weather Review, 143, 5115–5133, 2015.

Houtekamer, P. and Zhang, F.: Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation, Monthly Weather Review, 144, 4489–4532, 2016.

Houtekamer, P., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L.: A system simulation approach to ensemble prediction, Monthly Weather Review, 124, 1225–1242, 1996.

Houtekamer, P., Deng, X., Mitchell, H. L., Baek, S.-J., and Gagnon, N.: Higher resolution in an operational ensemble Kalman filter, Monthly Weather Review, 142, 1143–1162, 2014.

Houtekamer, P. L. and Mitchell, H. L.: Data assimilation using an ensemble Kalman filter technique, Monthly Weather Review, 126, 796–811, 1998.

Houtekamer, P. L. and Mitchell, H. L.: Ensemble kalman filtering, Quarterly Journal of the Royal Meteorological Society, 131, 3269–3289, 2005.

Kondo, K. and Miyoshi, T.: Impact of removing covariance localization in an ensemble Kalman filter: experiments with 10,240 members using an intermediate AGCM, Monthly Weather Review, 2016.

Laprise, R., De Elia, R., Caya, D., Biner, S., Lucas-Picher, P., Diaconescu, E., Leduc, M., Alexandru, A., Separovic, L., et al.: Challenging some tenets of regional climate modelling, Meteorology and Atmospheric Physics, 100, 3–22, 2008.

Leith, C.: Theoretical skill of Monte Carlo forecasts, Monthly Weather Review, 102, 409–418, 1974.

Luo, Y. and Chen, Y.: Investigation of the predictability and physical mechanisms of an extreme-rainfall-producing mesoscale convective system along the Meiyu front in East China: An ensemble approach, Journal of Geophysical Research: Atmospheres, 120, 2015.

Ménétrier, B., Montmerle, T., Berre, L., and Michel, Y.: Estimation and diagnosis of heterogeneous flow-dependent background-error covariances at the convective scale using either large or small ensembles, Quarterly Journal of the Royal Meteorological Society, 140, 2050–2061, 2014.

Migliorini, S., Dixon, M., Bannister, R., and Ballard, S.: Ensemble prediction for nowcasting with a convection-permitting model – I: description of the system and the impact of radar-derived surface precipitation rates, Tellus A, 63, 468–496, 2011.

Miyoshi, T., Kondo, K., and Imamura, T.: The 10,240-member ensemble Kalman filtering with an intermediate AGCM, Geophysical Research Letters, 41, 5264–5271, 2014.

Mullen, S. L. and Buizza, R.: The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system, Weather and Forecasting, 17, 173–191, 2002.

Penland, C.: Some Issues in Stochastic Weather/Climate Modeling or How do I use Stochastic Differential Equations to Model Something Real., pp. 87–100, 2011.

Pereira, M. B. and Berre, L.: The use of an ensemble approach to study the background error covariances in a global NWP model, Monthly Weather Review, 134, 2466–2489, 2006.

Pielke, R.: Mesoscale Meteorological Modeling, Academic Press, 2001.

Raynaud, L. and Bouttier, F.: Comparison of initial perturbation methods for ensemble prediction at convective scale, Quarterly Journal of the Royal Meteorological Society, 142, 854–866, 2016.

Raynaud, L., Berre, L., and Desroziers, G.: Objective filtering of ensemble-based background-error variances, Quarterly Journal of the Royal Meteorological Society, 135, 1177–1199, 2009.

Schraff, C., Reich, H., Rhodin, A., Schomburg, A., Stephan, K., Periáñez, A., and Potthast, R.: Kilometre-scale ensemble data assimilation for the COSMO model (KENDA), Quarterly Journal of the Royal Meteorological Society, 142, 1453–1472, 2016.

Schwartz, C. S. and Liu, Z.: Convection-permitting forecasts initialized with continuously cycling limited-area 3DVar, ensemble Kalman filter, and 'hybrid' variational–ensemble data assimilation systems, Monthly Weather Review, 142, 716–738, 2014.

Schwartz, C. S., Romine, G. S., Smith, K. R., and Weisman, M. L.: Characterizing and optimizing precipitation forecasts from a convection-permitting ensemble initialized by a mesoscale ensemble Kalman filter, Weather and Forecasting, 29, 1295–1318, 2014.

Schwartz, C. S., Romine, G. S., Weisman, M. L., Sobash, R. A., Fossell, K. R., Manning, K. W., and Trier, S. B.: A real-time convection-allowing ensemble prediction system initialized by mesoscale ensemble Kalman filter analyses, Weather and Forecasting, 30, 1158–1181, 2015.

Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, Monthly Weather Review, 139, 976–991, 2011.

Skamarock, W. C.: Evaluating mesoscale NWP models using kinetic energy spectra, Monthly Weather Review, 132, 3019–3032, 2004.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, paper presented at ECMWF Workshop on Predictability, Eur. Cent. for Med. Range Weather Forecasts, Reading, UK, pp. 1–25, 1997.

Tennant, W.: Improving initial condition perturbations for MOGREPS-UK, Quarterly Journal of the Royal Meteorological Society, 141, 2324–2336, 2015.

Tong, M. and Xue, M.: Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments, Monthly Weather Review, 133, 1789–1807, 2005.

Toth, Z. and Kalnay, E.: Ensemble forecasting at NCEP and the breeding method, Monthly Weather Review, 125, 3297–3319, 1997.

Vaughan, G., Methven, J., Anderson, D., Antonescu, B., Baker, L., Baker, T., Ballard, S., Bower, K., Brown, P., Chagnon, J., et al.: Cloud banding and winds in intense European cyclones: Results from the DIAMET project, Bulletin of the American Meteorological Society, 96, 249–265, 2015.

Wang, X. and Bishop, C. H.: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes, Journal of the Atmospheric Sciences, 60, 1140–1158, 2003.

Yashiro, H., Terasaki, K., Miyoshi, T., and Tomita, H.: Performance evaluation of a throughput-aware framework for ensemble data assimilation: the case of NICAM-LETKF, Geoscientific Model Development, 9, 2293–2300, 2016.

Zhang, F.: Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts, Monthly Weather Review, 133, 2876–2893, 2005.