The Cyclogenesis Butterfly: Uncertainty growth and forecast reliability during extratropical cyclogenesis

Mark J. Rodwell[1] and Heini Wernli[2]

[1]European Centre for Medium–Range Weather Forecasts, Reading, UK
[2]Institute for Atmospheric and Climate Science, ETH Zürich, Switzerland

Correspondence: Mark Rodwell (mark.rodwell@ecmwf.int)

Abstract.

The chaotic and multi-scale nature of the atmosphere was brought into the public imagination by the question of Lorenz in 1972: "Does the flap of a butterfly's wings in Brazil set off a tornado in Texas?". While numerical models currently used in global weather prediction have grids which certainly do not resolve butterflies, they nevertheless display strong sensitivity to initial conditions. In ensemble forecasts, this sensitivity is manifested in the growth of ensemble variance with leadtime. Interestingly, in the extratropics, this uncertainty growth appears to be organised in particular synoptic flow configurations which, for the purposes of practical prediction, might be thought of as large, metaphorical 'butterflies' in the flow. Often, forecasts for severe downstream weather show marked improvement in skill when these 'butterflies' have passed. Here we focus on the "Cyclogenesis Butterfly" — associated with baroclinic and convective instabilities in the extratropics. We investigate four operational ensemble forecast systems within the TIGGE archive, and find that they display quite different initial uncertainty growth rates in cases of cyclogenesis. Evaluation through use of an extended spread–error equation shows that some models fail to maintain short-range statistical reliability within the North Atlantic stormtrack during winter 2020/21. For the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, flow-type clustering demonstrates that its 'over-spread' in the stormtrack is indeed associated with cyclogenesis events. At day 2, a large part of the total spread in cyclogenesis cases is associated with the growth of initial uncertainty (as derived from ensemble data assimilation), but up to 25% can be associated with the additional singular vector perturbations to the initial conditions, and up to 25% with the representation of model uncertainty. The sensitivities of spread to resolution, the explicit representation of convection, and the assimilation of local observations are also considered. The study raises the question whether a reduction in singular vector perturbations, to improve stormtrack reliability, would then allow short-range diagnostics to better inform further model and model-uncertainty development, which could be beneficial throughout the forecast range.

1 Introduction

The chaotic nature of the atmosphere, with its large sensitivity to small perturbations (Sutton, 1954; Lorenz, 1963, 1972), poses a challenge for numerical weather prediction (NWP). To embrace this challenge, NWP has developed ensemble techniques, whereby a set (or ensemble) of initial conditions, representing the uncertainty in the current state, is used to initialise an

ensemble of forecasts (Palmer et al., 1992; Molteni et al., 1996). The variance of the ensemble, which generally grows with leadtime, is then a measure of forecast uncertainty — something of great interest to the user.

This study focuses on uncertainty growth in the North American / North Atlantic / European region, and particularly the North Atlantic winter stormtrack, with its embedded cyclogenesis events and other synoptic systems. Synoptic scales are particularly important in NWP because these become the largest contributor to global ensemble variance over the first few days of the forecast (Tribbia and Baumhefner, 2004). Previous studies have shown that occasional drops in weather forecast performance in the region of interest can be related to particular prior synoptic flow situations. These include baroclinic flows propone to cyclogenesis (Lillo and Parsons, 2017), situations of convective instability (Rodwell et al., 2013; Sun and Zhang, 2016), the extratropical transition of tropical cyclones (Riemer and Jones, 2014), development of cut-off vortices, and their interaction with upper-tropospheric troughs (Grams et al., 2018; Baumgart and Riemer, 2019). The impacts of these events on jet stream/waveguide dynamics can lead to errors and uncertainties in downstream 'blocking' (Rodwell et al., 2013) and extreme precipitation (Grams and Blumer, 2015). A 'Lagrangian growth rate' diagnostic (Rodwell et al., 2018) highlights these flow configurations as situations of enhanced 12 h growth in synoptic-scale ensemble spread in upper-tropospheric potential vorticity. The possibility that this growth rate diagnostic provides a useful means of systematically studying the flow-dependence of forecast skill and reliability motivates the present study. Here, the focus is on the strong growth rates associated with extratropical cyclogenesis — hence the term 'The Cyclogenesis Butterfly'. It should be emphasised that this term refers to the sensitivity to initial uncertainty derived from operational ensemble data assimilation (Isaksen et al., 2010), rather than intrinsic growth rates associated with the atmosphere's sensitivity to small scale perturbations (Lorenz, 1963, 1972). See, e.g., Durran and Gingrich (2014) and Palmer et al. (2014) for interesting discussions. Although the underlying processes driving growth (Baumgart et al., 2019) are likely to be similar, global operational ensembles generally require a representation of 'model uncertainty' (MU; Buizza et al., 1999) — which aims to account for missing interactions with scales that are unresolved on the model grid — in order to achieve forecast 'reliability' (Gneiting and Raftery, 2007; Rodwell et al., 2020).

While there are clear differences with intrinsic growth rates (Selz et al., 2022), a useful comparison might be made between different operational ensemble systems — for example those available in near real-time from 'The International Grand Global Ensemble' (TIGGE, Swinbank et al., 2016) archive. A question addressed here is "How similar are the growth rates in such ensembles in cases of extratropical cyclogenesis?" A means of evaluating each system's short-range 'reliability' (Gneiting and Raftery, 2007; Rodwell et al., 2020) would be useful in this comparison. At short-ranges, the 'spread–error' reliability equation (Leutbecher and Palmer, 2008, which states that the mean ensemble variance should agree with the mean-squared-error of the ensemble mean) needs to be extended to account for biases and non-negligible uncertainties in the verifying analysis. When combined with flow-type clustering, the approach should allow a flow-dependent evaluation of short-range reliability and initial growth rates — something this study attempts to do. These could depend on, for example, the reliability of initial conditions (Rodwell et al., 2016), whether an ensemble system employs singular vector (SV; Molteni and Palmer, 1993) perturbations to boost initial growth rates, how convection is parametrized or resolved (Wedi et al., 2020) and, as discussed above, how MU is represented. In the absence of SV perturbations, the potential is raised that developments to the model and MU, which improve flow-dependent short-range reliability, will be beneficial throughout the forecast range.

The study is structured as follows. Section 2 lists the data sources used. (There is a series of methodologies employed in the study — these are discussed later, once the motivation for their use becomes apparent). Section 3 discusses the quantification and interpretation of uncertainty growth rates, with particular application to cyclogenesis. Section 4 discusses forecast reliability in the presence of bias and uncertainty in the verifying analyses, with application to seasonal means in the North Atlantic stormtrack region, and when composited on cyclogenesis cases. Section 5 investigates the sources of uncertainty in cyclogenesis cases. Conclusions and a discussion about prospective research are given in Sect. 6. Supplementary material includes animations of uncertainty growth rates.

2 Models, data sources and key parameters

2.1 The ECMWF ensemble assimilation and forecast system

The underlying earth system model for the ECMWF Ensemble of Data Assimilations (EDA; Isaksen et al., 2010), and Ensemble forecast (ENS; Palmer et al., 1992; Molteni et al., 1996) uses spherical harmonics to compute much of the dynamics, with physical parametrizations computed in grid-point space. Of particular interest here is the parametrization of convection, which was originally based on Tiedtke (1989), but includes revisions to entrainment and coupling with the large scale (Bechtold et al., 2008; Hirons et al., 2013), and improvements in the diurnal cycle of convection through use of a modified convective available potential energy (CAPE) closure (Bechtold et al., 2014).

The 50-member EDA used here has a nominal horizontal grid resolution of ~ 16 km. More specifically, in the final EDA iteration, the underlying non-linear model retains a "Triangular" configuration of spherical harmonics with total wavenumbers $\leqslant 639$ and uses a "Cubic octahedral" grid so that 4 grid points represent the smallest waves (the resolution is thus summarised as TCo639), has 137 levels in the vertical (L137) and uses a 12 min timestep. 4D Variational data assimilation (4DVar, Rabier et al., 2000) uses the full non-linear model, together with 'tangent-linear' and adjoint versions, to extract the information content from many millions of conventional and satellite observations during each 12 h analysis cycle. This is done by first screening observations and correcting them using Variational Bias Correction (VarBC, Dee, 2004). For each EDA member, the observations are then randomly perturbed to simulate observation uncertainty (Isaksen et al., 2010). The 'background' for a given EDA member is the non-linear forecast initialised from the same member's previous analysis. The estimated uncertainty in this forecast is based on the variances and correlations between the ensemble of background forecasts, with a climatological contribution to correlations for improved stability. For each EDA member, 4DVar combines its background forecast and perturbed observation set in a way that is consistent with the estimated uncertainties in the background (Bonavita et al., 2016) and observations (Geer et al., 2018). An additional unperturbed EDA member, with no perturbations to the observations, is also made.

The EDA is used to initiate a 50-member ENS, with resolution TCo639, L91 and a 12 min timestep. Initial conditions are re-centred on a more recent ("Early Delivery") unperturbed high-resolution (HRES) 4DVar analysis. 'Singular Vector' (SV; Molteni and Palmer, 1993; Leutbecher and Lang, 2014) perturbations are added to the initial conditions as a pragmatic means of boosting ENS spread over the first 2 days. A model uncertainty (MU) parametrization, which partly aims to represent scale

Table 1. Details of the four TIGGE models used in this study, valid during the period of investigation. EDA=Ensemble of 4DVar (Isaksen et al., 2010), 4DVar=4D Variational data assimilation (Rabier et al., 2000), EnKF=Ensemble Kalman Filter (Evensen, 1994), ETKF=Ensemble Transform Kalman Filter (Bishop et al., 2001), LETKF=Local Ensemble Transform Kalman Filter (Hunt et al., 2007), SPPT=Stochastic Perturbation to Physical Tendencies (Buizza et al., 1999), SV=Singular Vector (Molteni and Palmer, 1993), RP=Random Parameters (McCabe et al., 2016), SKEB=Stochastic Kinetic Energy Backscatter (Shutts, 2004).

Centre	ECMWF	JMA	NCEP	UKMO
Resolution (nominal)	16 km	42 km	25 km	21 km
Vertical levels	91	100	64	70
Number of perturbed members	50	26	30	17
Run times (UTC)	0,12	00,12	00,06,12,18	00,06,12,18
Initial perturbation strategy	EDA, SV	LETKF, SV	EnKF	ETKF
Model uncertainty representation	SPPT	SPPT	SPPT, SKEB	RP, SKEB

interactions with (missing) sub-grid-scale variations, is important for the general growth of ENS spread into the medium-range (e.g., to day 10). Here, this MU representation is based on "Stochastic Perturbation to Physical Tendencies" (SPPT, Buizza et al., 1999), which represents a perturbation to the *total* physical tendency, and is applied throughout the forecast range. Note that SPPT is also applied to the non-linear model in the perturbed members of the EDA, but there are no SV perturbations in the EDA.

The sensitivity experiments discussed in Sect. 5 are based on cycle 46r1 of the ECMWF Integrated Forecasting System (IFS). This cycle was operational from 11 June 2019 to 20 June 2020. In these experiments, VarBC and SPPT are "warm-started" from the operational assimilation, ENS initial condition re-centring is not done, and SV perturbation scaling is based on the current EDA cycle rather than the previous cycle. In the higher-resolution ~ 4 km ENS experiment, the model is run in "single precision" (Váňa et al., 2017; Lang et al., 2021) with resolution TCo2559 L91 and a 4 min timestep.

The study additionally makes use of ECMWF reanalysis version 5 (ERA5; Hersbach et al., 2020), which is also based on the EDA.

2.2 Other ensemble forecast systems in the TIGGE archive

In comparisons with ensemble forecasts from other centres, all data are retrieved from the TIGGE archive, which currently contains global ensemble forecasts from about a dozen of the world's operational forecasting centres (12 centres were present on 1 September 2021). These forecasts are available a few days behind real-time and are a valuable resource for diagnostic studies. Three other TIGGE models are compared with the ECMWF model. These are from the Japan Meteorological Agency (JMA), the United States' National Centers for Environmental Prediction (NCEP), and the United Kingdom's Met Office (UKMO). Salient details of these four ensemble forecast systems are presented in Table 1. For consistency, comparisons are based on 00 and 12 UTC run times only.

2.3 Potential vorticity

A useful quantity for this study is isentropic potential vorticity (IPV, Hoskins et al., 1985), $P = -g(f + \zeta_\theta)\frac{\partial \theta}{\partial p}$. Here, g is the gravitational acceleration, f is the Coriolis parameter, p is pressure, θ is potential temperature, and $\zeta_\theta = \boldsymbol{k} \cdot \boldsymbol{\nabla}_\theta \times \boldsymbol{v}$ is the isentropic vorticity, where $\boldsymbol{k}$ is the local unit vertical vector, $\boldsymbol{\nabla}_\theta$ is the horizontal gradient operator on an isentropic surface, and $\boldsymbol{v}$ is the horizontal wind vector. IPV is usually measured in PV units (PVU) with $1\,\mathrm{PVU} = 10^{-6}\,\mathrm{m^2\,s^{-1}K\,kg^{-1}}$. A key advantage of using IPV here is that its tendencies due to dynamic and diabatic effects can be readily disentangled:

$$\partial P/\partial t + \boldsymbol{v} \cdot \boldsymbol{\nabla}_\theta P = \mathcal{D} \quad , \tag{1}$$

where $\mathcal{D}$ represents the effects of non-conservative (diabatic and frictional) processes (Holton, 2004, Eq. 4.36). IPV is thus conserved following the horizontal flow on an isentrope in the absence of such processes. This study will use $P315$, the IPV on the 315 K isentrope, which typically intersects the dynamical tropopause (P=2 PVU) during winter in the midlatitudes.

3 Cyclogenesis and the growth of uncertainty

3.1 An example of cyclogenesis

Before performing a more systematic investigation of the 'cyclogenesis butterfly', a specific case of North Atlantic cyclogenesis is selected to illustrate and highlight the salient features in its development and its uncertainty. The case was chosen because it was quite "clean", without being strongly affected by other flow perturbations in its environment, and because it involved both baroclinic and diabatic aspects. However, the speed of cyclogenesis and growth of uncertainty were not considered in the choice. Figure 1 introduces the main synoptic-scale flow features of the event. The cyclone develops on 26 November 2019 over eastern North America (not shown). A day later, it reaches the Great Lakes with a minimum pressure at mean sea level (PMSL) of about 990 hPa (not shown). At 18 UTC on 28 November, the cyclone reaches the North American east coast with a similar intensity (Fig. 1a). There is a cutoff in potential vorticity on the $\theta = 315$ K isentrope ($P315$) aloft, and intense surface rainfall in the region identified as a warm conveyor belt (WCB). One day later (Fig. 1b), the cyclone has strongly deepened to below 974 hPa as it moves slowly out into the Atlantic. Intense precipitation continues in the WCB ascent regions along the cold and bent-back fronts. In the next 24 h, the cyclone remains stationary, and it deepens further to below 970 hPa (Fig. 1c). The WCB and associated band of intense precipitation now extend from about 25 to 55°N, and the $P315$ pattern attains the classical structure of a mature cyclone, with a large-amplitude trough-ridge dipole up- and downstream of the surface cyclone, respectively. The heterogeneity of the precipitation rate along the WCB is reminiscent of the occurrence of embedded convection (Oertel et al., 2020). This brief overview emphasizes the strong deepening of the system off the North American east coast and its combined baroclinic and diabatic character. Forecast experiments that investigate the strong cyclone deepening period for this case are initialised at 12 UTC on 28 November 2019, and therefore the dates shown in Fig. 1 correspond to forecast days 0, 1 and 2.

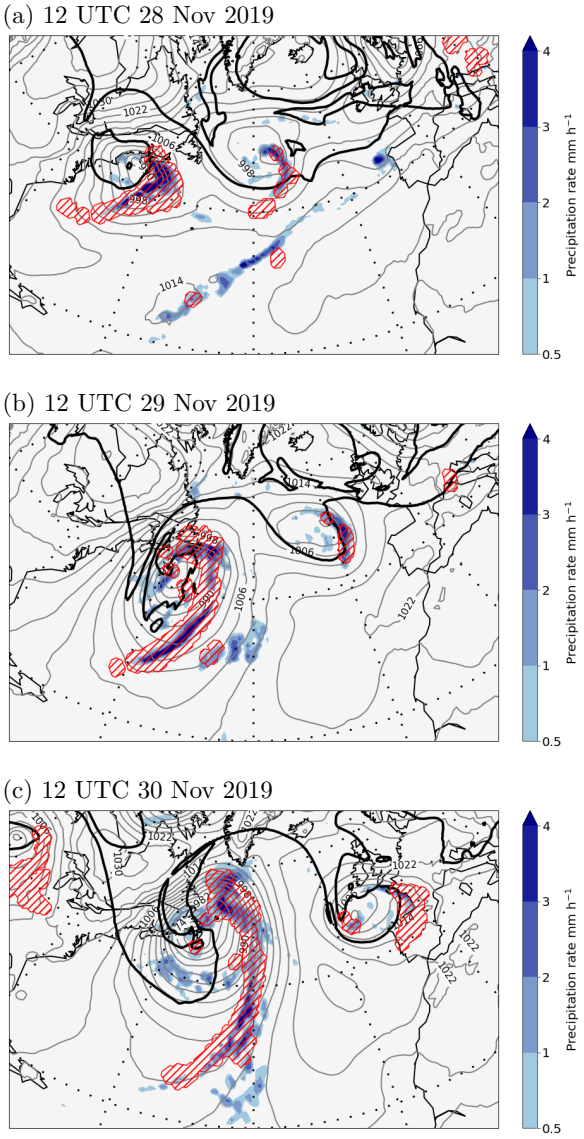

Figure 1. Synoptic overview, based on ERA5 reanalyses, of a North Atlantic cyclogenesis event at 12 UTC on (a) 28 November, (b) 29 November, and (c) 30 November 2019. Shown are PMSL (grey contours at intervals of 4 hPa), surface precipitation accumulated over the previous hour (colour shading, in mm h^{-1}), the $P315 = 2$ PVU contour indicating the tropopause on 315 K (black solid line), and the WCB ascent region (red hatching). WCBs are identified as 48 h trajectories that ascend from the lower troposphere by more than 600 hPa in 48 h (e.g., Wernli and Davies, 1997; Madonna et al., 2014), and the ascent region (shown with red hatching) corresponds to the envelope of the horizontal positions of all WCB trajectories, initialized every 6 h, that are located between 800 and 500 hPa at the indicated time (corresponding approximately to their mid-ascent time).

3.2 The Lagrangian growth rate

The local uncertainty growth rate of an ensemble forecast can be estimated as $\widehat{\sigma}_X^{-1}\,\partial\widehat{\sigma}_X/\partial t$ where $\widehat{\sigma}_X$ is the ensemble standard deviation of some atmospheric parameter field X (the circumflex $\widehat{\ }$ indicates a sample estimate throughout this study) and t is the forecast leadtime. As observed by Rodwell et al. (2018), a large component of the local growth rate can be associated with the advection of uncertainty, and it is useful to separate this off to highlight the processes driving uncertainty growth. If X is chosen to be IPV P, then Eq. (1) provides a straightforward approach to achieve this. Firstly, it is useful to establish some notation. For any parameter Y (e.g., P, $\boldsymbol{v}$, $\mathcal{D}$), write Y_i for its value in ensemble forecast member $i \in \{1,\ldots,m\}$. Using an overline to denote a mean $\overline{Y} = m^{-1}\sum_i Y_i$ (even for non-linear terms: $\overline{YZ} = m^{-1}\sum_i Y_i Z_i$) and a prime to denote deviations from the mean $Y_i' = Y_i - \overline{Y}$, then the variance estimator for P can be written as $\widehat{\sigma}_P^2 = \overline{P'^2}$. Using this notation and the results that $\overline{Y'} = 0$ and $\overline{Y'\overline{Z}} = 0$, the local growth rate for P can be written as

$$
\begin{aligned}
\frac{1}{\widehat{\sigma}_P}\frac{\partial\widehat{\sigma}_P}{\partial t} &= \frac{1}{2\widehat{\sigma}_P^2}\frac{\partial\widehat{\sigma}_P^2}{\partial t} \\
&= \frac{1}{2\widehat{\sigma}_P^2}\frac{\partial\overline{P'^2}}{\partial t} \\
&= \frac{1}{\widehat{\sigma}_P^2}\overline{P'\left(\frac{\partial P}{\partial t} - \frac{\partial\overline{P}}{\partial t}\right)} \\
&= \frac{1}{\widehat{\sigma}_P^2}\overline{P'\left(\mathcal{D} - \boldsymbol{v}\cdot\boldsymbol{\nabla}_\theta P\right)} \\
&= \frac{1}{\widehat{\sigma}_P^2}\overline{P'\left(\mathcal{D}' - \overline{\boldsymbol{v}}\cdot\boldsymbol{\nabla}_\theta P' - \boldsymbol{v}'\cdot\boldsymbol{\nabla}_\theta P\right)} \\
&= -\frac{1}{2\widehat{\sigma}_P^2}\overline{\boldsymbol{v}}\cdot\boldsymbol{\nabla}_\theta\overline{P'^2} + \frac{1}{\widehat{\sigma}_P^2}\overline{P'\left(\mathcal{D}' - \boldsymbol{v}'\cdot\boldsymbol{\nabla}_\theta P\right)} \\
&= -\frac{1}{\widehat{\sigma}_P}\overline{\boldsymbol{v}}\cdot\boldsymbol{\nabla}_\theta\widehat{\sigma}_P + \frac{1}{\widehat{\sigma}_P^2}\overline{P'\mathcal{D}'} - \frac{1}{\widehat{\sigma}_P^2}\overline{P'\boldsymbol{v}'\cdot\boldsymbol{\nabla}_\theta P} \quad ,
\end{aligned}
\tag{2}
$$

This is a different formulation of the equation explored by Baumgart and Riemer (2019); their equation (8). The first term on the last line is the advection of uncertainty by the ensemble mean wind. Rearranging, we can define a 'Lagrangian growth rate' for IPV as

$$
\mathrm{LGR}_P \equiv \frac{1}{\widehat{\sigma}_P}\left\{\frac{\partial\widehat{\sigma}_P}{\partial t} + \overline{\boldsymbol{v}}\cdot\boldsymbol{\nabla}_\theta\widehat{\sigma}_P\right\} = \frac{1}{\widehat{\sigma}_P^2}\overline{P'\mathcal{D}'} - \frac{1}{\widehat{\sigma}_P^2}\overline{P'\boldsymbol{v}'\cdot\boldsymbol{\nabla}_\theta P} \quad . \tag{3}
$$

The left side of Eq. (3), as in Rodwell et al. (2018), represents the rate of growth of the ensemble standard deviation of P following the ensemble mean horizontal flow on an isentrope. The two terms on the right hand side of Eq. (3) provide a useful glimpse at the interactions that drive this growth. The second term on the right is the covariance between P' and the advection of PV by each member's anomalous wind $\boldsymbol{v}'\cdot\boldsymbol{\nabla}_\theta P$. This term represents the effect of interactions between

dynamical uncertainties at all (resolved) scales. In the case of initial growth rates in operational ensemble forecasts, this could include interactions with large-scale analysis uncertainties, particularly positive contributions associated with cyclogenesis (*viz* Hoskins et al., 1985, their Fig. 21) since there is as much variance power in the EDA at scales ~ 5000 km as there is at scales ~ 50 km. (Maximum EDA variance contributions come from intermediate scales ~ 400 km — illustrative power spectra are presented in Fig. B1 in Appendix B). This initial large-scale contribution is likely to be in contrast to some predictability studies (Judt, 2018), which apply grid-point Gaussian noise and where the smallest resolved scales will dominate the total initial variance. In addition, if (dry) singular vector perturbations are applied to the initial conditions for the operational forecast (Table 1), then the growth associated with baroclinic uncertainties will be accentuated. The first term on the right hand side of Eq. (3) represents covariances between P' and non-conservative processes $\mathcal{D}'$. There is the posibility for negative growth rate contributions here, associated with the effects of latent heating on upper tropospheric $\mathcal{D}'$ during cyclogenesis (*viz* Ahmadi-Givi et al., 2004, their Fig. 14). Here too, there could be differences with intrinsic growth rates — for example, due to the need to parametrize turbulent processes, and because the model uncertainty representation can only approximate the effects of interactions with sub-grid-scale variations. In operational and intrinsic contexts, as the leadtime increases and smaller scale uncertainties become saturated, the larger scales will become increasingly important for driving further growth. Hence there is no fixed 'ground-truth' with which to evaluate modelled growth rates. Nevertheless, it is informative to calculate operational growth rates, and to compare these for different models.

In Sect. 3.3, the growth rate on the left side of Eq. (3) will be calculated from the background forecasts of the EDA. The right side of Eq. (3) is discussed further in relation to prospective research in Sect. 6.

3.3 Uncertainty growth in the EDA

This study is interested in the growth of synoptic scale uncertainty (due to interactions between all scales) during the first few days of the ensemble forecast. To focus on this growth, LGR_P is filtered with a synoptic spatio-temporal filter. This multiplies spectral coefficients with total wavenumber $n > n_s = 21$ by $\{n_s(n_s+1)\}/\{n(n+1)\}$ so that scales larger than ~ 700 km are retained. The filter also includes a 24 h running-mean. To understand how growth rates depend on the synoptic flow situation, it is useful to consider very short leadtimes when all ensemble members are representing essentially the same synoptic flow situation. A natural choice is to use the short background forecasts from ensemble data assimilation.

Figure 2 shows (shaded) the filtered LGR_P for $P315$ based on the EDA 12 h background forecasts, centred at 12 UTC on 29 November 2019 (one day into the strong cyclogenesis period shown in Fig. 1). The caption to Fig. 2 provides technical details of the calculation of LGR_P. Also shown in Fig. 2, from the unperturbed EDA member, are $P315 = 2$ PVU (red contour) and vectors display the 850 hPa horizontal wind ($\mathbf{v}850$, coloured blue when the 850 hPa horizontal moisture flux exceeds 100 $\mathrm{g\,kg^{-1}\,m\,s^{-1}}$). Black dots indicate the ensemble-mean precipitation rate. Large LGR_P values are seen at the southern extent of the upper-level trough. Note that the orange contours at the centre of this region have the same interval as that of the shading, and indicate values in excess of $0.18\,\mathrm{h^{-1}}$. This location of large LGR_P values appears consistent with Hoskins et al. (1985), discussed above, but further investigation would be useful. Also evident in Fig. 2 are weaker negative LGR_P values (where the ensemble is tending to converge in this Lagrangian sense) particularly in the ridge-building region above the northern part of

the WCB (see Fig. 1b). Further investigation could help confirm if this is associated with the erosion of the eastern edge of the upper level trough (Ahmadi-Givi et al., 2004), discussed above.

The supplementary material includes animations of similar plots to Fig. 2. They show fields of $P315$, $\mathbf{v}850$ and precipitation which effectively 'shadow' the true synoptic evolution of the flow, with LGR_P highlighting the initial (~ 12 h) rate of divergence of the ensemble about the synoptic state. Synoptic features associated with large model growth rates are evident. The results in Fig. 2 are typical of many cases seen within the animations. In addition to these cyclogenesis cases, which often include strongly precipitating WCBs (with embedded convection, Oertel et al., 2020), other common situations for strong growth rates (not investigated here, but consistent with previous studies) are during the extratropical transitions of tropical cyclones (Riemer and Jones, 2014), and within high CAPE situations over North America where mesoscale convection is likely to develop (Rodwell et al., 2013; Sun and Zhang, 2016; Rodwell et al., 2018). All of these situations can lead to deterministic forecast 'busts' or 'dropouts' (Lillo and Parsons, 2017), extreme precipitation (Grams and Blumer, 2015) or blocking events (Rodwell et al., 2013) over Europe.

3.4 Uncertainty growth in TIGGE forecasts

As discussed in Sect. 3.2, there are reasons why the growth rates displayed by the ECMWF model within its operational EDA might not agree with intrinsic growth rates associated with the atmosphere's sensitivity to small scale perturbations. The question arises as to how well the ECMWF growth rates agree with the growth rates derived from the ensembles of other operational forecast centres? This question is explored using the first 12 h of ensemble forecasts within the TIGGE archive, for the models summarised in Sect. 2.2. Since potential vorticity is not available in TIGGE, the pragmatic decision is made to calculate the Lagrangian growth rate $\mathrm{LGR}_Z = \widehat{\sigma}_Z^{-1}(\partial\widehat{\sigma}_Z/\partial t + \overline{\boldsymbol{v}}\cdot\boldsymbol{\nabla}_p\widehat{\sigma}_Z)$ for $Z = Z250$, the geopotential height field at 250 hPa, where $\boldsymbol{\nabla}_p$ is the horizontal gradient operator on the pressure surface.

Figure 3 shows filtered LGR_Z (shaded) for the TIGGE models centred on the same time as that shown in Fig. 2. The caption to Fig. 3 provides technical details of the calculation of LGR_Z. Other fields shown are the same as in Fig. 2 except that $Z250$ is contoured in green. For ECMWF in the region of cyclogenesis, filtered LGR_Z from the ENS (Fig. 3a) agrees quantitatively quite well with filtered LGR_P from the EDA (Fig. 2), despite being growth rates of different fields, and despite the use of additional singular vector perturbations in the ENS. One difference of note for later might be that the maximum ENS growth rate is placed a little more towards the western side of the upper level trough than is the case for the EDA growth rate (cf Fig. 2, Fig. 3a).

Comparison amongst the models in Fig. 3 indicates strong agreement, over the first 12 h, in their portrayal of the synoptic situation, as displayed in the fields of $Z250$, $\mathbf{v}850$, moisture fluxes and, to some extent, precipitation. Although there are commonalities, such as in the observational information available to each centre's data assimilation, this agreement suggests that the short range forecasts of all models shadow well the true synoptic evolution. Despite this agreement, the filtered LGR_Z differs widely amongst the models in this example. Looking over many examples (within the TIGGE animation for the DJF 2020/21 season in the supplementary material), the agreement can be better (Fig. 4). Nevertheless, differences between the models' growth rates can be striking, with the ECMWF model tending to display the strongest values. The question arises as

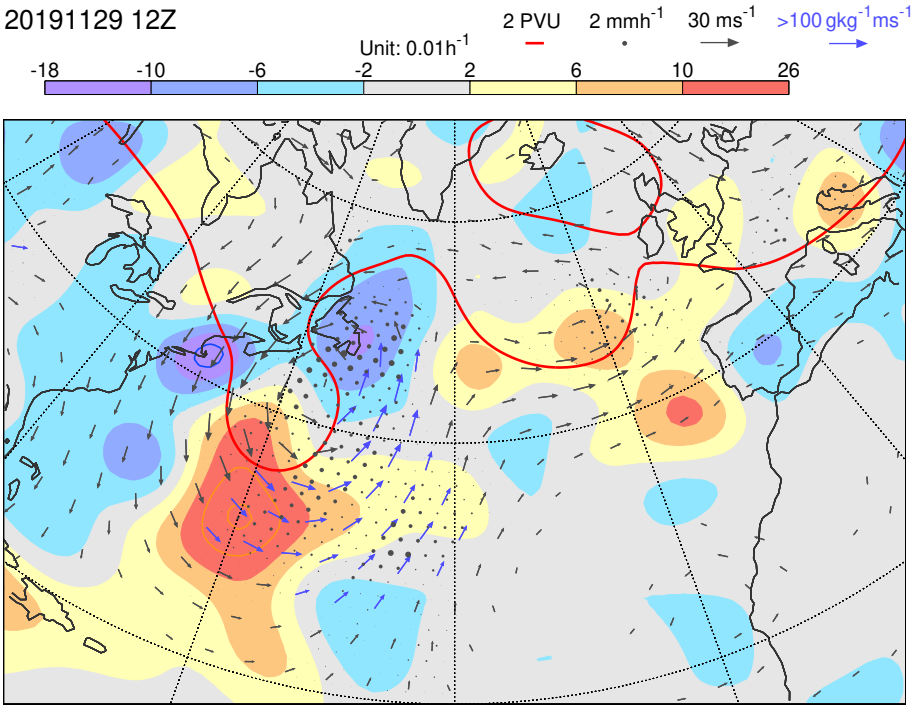

Figure 2. Growth rate LGR_P for $P315$ in EDA background forecasts (shaded), centred at 12 UTC on 29 November 2019. Note that orange and blue contours extend the shading scheme, with the same interval. In these cases, the most extreme values are indicated at the ends of the colour bar. Also shown, for the unperturbed EDA member, are $P315 = 2$ PVU (red contour) and $\mathbf{v}850$ (vectors, which are plotted blue if the moisture flux at 850 hPa exceeds 100 $\mathrm{g\,kg^{-1}\,m\,s^{-1}}$). The ensemble mean precipitation is indicated with black dots (with radius proportional to the precipitation rate up to the maximum size at 2 $\mathrm{mmh^{-1}}$). The fields shown are constructed using the 12 h background forecasts from the EDA (so no lead-times are greater than 12 h), started at 06 and 18 UTC. Calculations use centred-means and differences between consecutive hourly leadtimes. For the upper tropospheric fields, $P315$ and zonal and meridional winds at $\theta = 315$ K ($u315$ and $v315$) are first interpolated to an "N32" reduced Gaussian grid (with 32 latitudes between the pole and equator). The spatial derivatives within the advection term in LGR_P are calculated using spectral transforms to and from a "T42" spherical harmonic representation. Note that N32 is sufficient to avoid aliasing of higher harmonics of the quadratic advection term into the T42 representation. The 12 fields of $P315$ and LGR_P are then concatenated over EDA cycles and spatially smoothed with a synoptic spatio-temporal filter (details in main text). The resulting timeseries of fields can be used to produce animations of $P315$ which 'shadow' the true synoptic evolution of the flow, with LGR_P highlighting the initial (~ 12 h) rate of divergence of the ensemble about the synoptic state. For the lower-tropospheric fields shown, zonal and meridional winds and specific humidities at 850 hPa ($u315$, $v315$, $q850$) and surface pressure p_*, all from the background forecasts of the control EDA member, are first interpolated to an "O32" octahedral reduced Gaussian grid (with 32 latitudes between the pole and equator). Values are set to "missing" where the 850 hPa surface is below the land surface (where $p_* < 850$ hPa) and the moisture flux is calculated as $q850\sqrt{u850^2 + v850^2}$. The ensemble mean total precipitation rate is used to indicate where precipitation is likely to occur. This is obtained on a higher resolution "O80" octahedral grid to give a good symbolic representation (stippling) of rainfall. After similar concatenation of EDA cycles, the lower-tropospheric fields are smoothed with a 24 h running mean.

to whether the ECMWF growth rates are too strong in the vicinity of cyclogenesis? Is the ECMWF ensemble spread being inflated unnecessarily in these situations, with the consequent impact on forecast scores?

The differences in the models' filtered LGR_z reflect differences in initialisation procedures, differences in the representation of model uncertainty, and differences in the deterministic models themselves (summarised in Table 1). While it is difficult to evaluate these growth rates per se, it is possible to assess how well each ensemble system maintains short-range statistical reliability within the North Atlantic stormtrack. This will be done in Sect. 4.3, after considering important conceptual aspects of short-range reliability and introducing a novel spread–error relationship in Sect. 4.1 and 4.2.

4 Forecast reliability

Reliability (Sanders, 1958) is a key attribute of ensemble forecasts (Gneiting and Raftery, 2007), allowing users to make unbiased decisions (Rodwell et al., 2020). The concept of reliability can be understood with reference to the schematic in Fig. 5. For clarity, the figure is shown in two dimensions and could relate, for example, to the prediction of the location of a storm. For a reliable forecast system (Hamill, 2001; Saetra et al., 2004), the verifying truth T should be statistically indistinguishable from any random sampling of the forecast distribution F, indicated by the grey concentric circles. This distribution, which need not be Gaussian or even uni-modal, has mean μ_F and standard deviation of distances from the mean σ_F (grey dashed line). Introducing a suffix j to indicate the forecast initiated at time t_j with $j \in \{1, \ldots, n\}$ then, since F_j is reliable (and assumed to be the only information available at time t_j about T_j), the expectation $\mathbb{E}_j[\cdot]$ at time t_j is that $\mathbb{E}_j[T_j] = \mu_{Fj}$ and $\mathbb{E}_j[(T_j - \mu_{Fj})^2] = \sigma^2_{Fj}$. The latter condition leads to the common 'spread–error' relation in ensemble forecasting — that reliability requires $\overline{(T - \widehat{\mu}_F)^2} \approx \overline{\widehat{\sigma}^2_F}$, where $\widehat{\mu}$ and $\widehat{\sigma}$ are the mean and standard deviation estimators, respectively, based on the sample of m ensemble members, and an overline indicates a mean of the n forecasts. Equality here can be improved by accounting for the finiteness of m and increasing n (Leutbecher and Palmer, 2008).

For the short forecast leadtimes of interest here, the spread–error relation is not adequate in general. This is because uncertainties in the knowledge of the verifying truth can be non-negligible compared to forecast variances at short leadtimes. Ensemble data assimilation (such as the EDA at ECMWF discussed in section 2.1) aims to estimate the uncertainty in the knowledge of the truth. A reliable analysis distribution A, which is consistent with the indicated truth T, is shown in Fig. 5 using blue concentric circles, with mean μ_A and standard deviation of distances from the mean σ_A (blue dashed line).

Let $\widetilde{F}$ and $\widetilde{A}$ be the underlying distributions associated with the *operational* forecast and verifying analysis. These distributions might not be reliable. For example, bias in these distributions could have a non-negligible impact on reliability — their means $\mu_{\widetilde{F}}$ and $\mu_{\widetilde{A}}$, respectively, are shown offset from those of the reliable distributions in Fig. 5. The ensemble (sample) means of these distributions are indicated by $\widehat{\mu}_{\widetilde{F}}$ and $\widehat{\mu}_{\widetilde{A}}$. These latter two parameters, along with their difference (or 'departure') d, are the only parameters shown in Fig. 5 that are ascertainable here, being obtained from the operational ensemble forecast system. Perhaps more importantly for reliability, the distributions $\widetilde{F}$ and $\widetilde{A}$ could have deficiencies in their variances $\sigma^2_{\widetilde{F}}$ and $\sigma^2_{\widetilde{A}}$, respectively. Are they under-spread or over-spread with respect to the variances of the reliable distributions, for example? Again, it is only the ensemble sample estimators $\widehat{\sigma}^2_{\widetilde{F}}$ and $\widehat{\sigma}_{\widetilde{A}}$ that are ascertainable here.

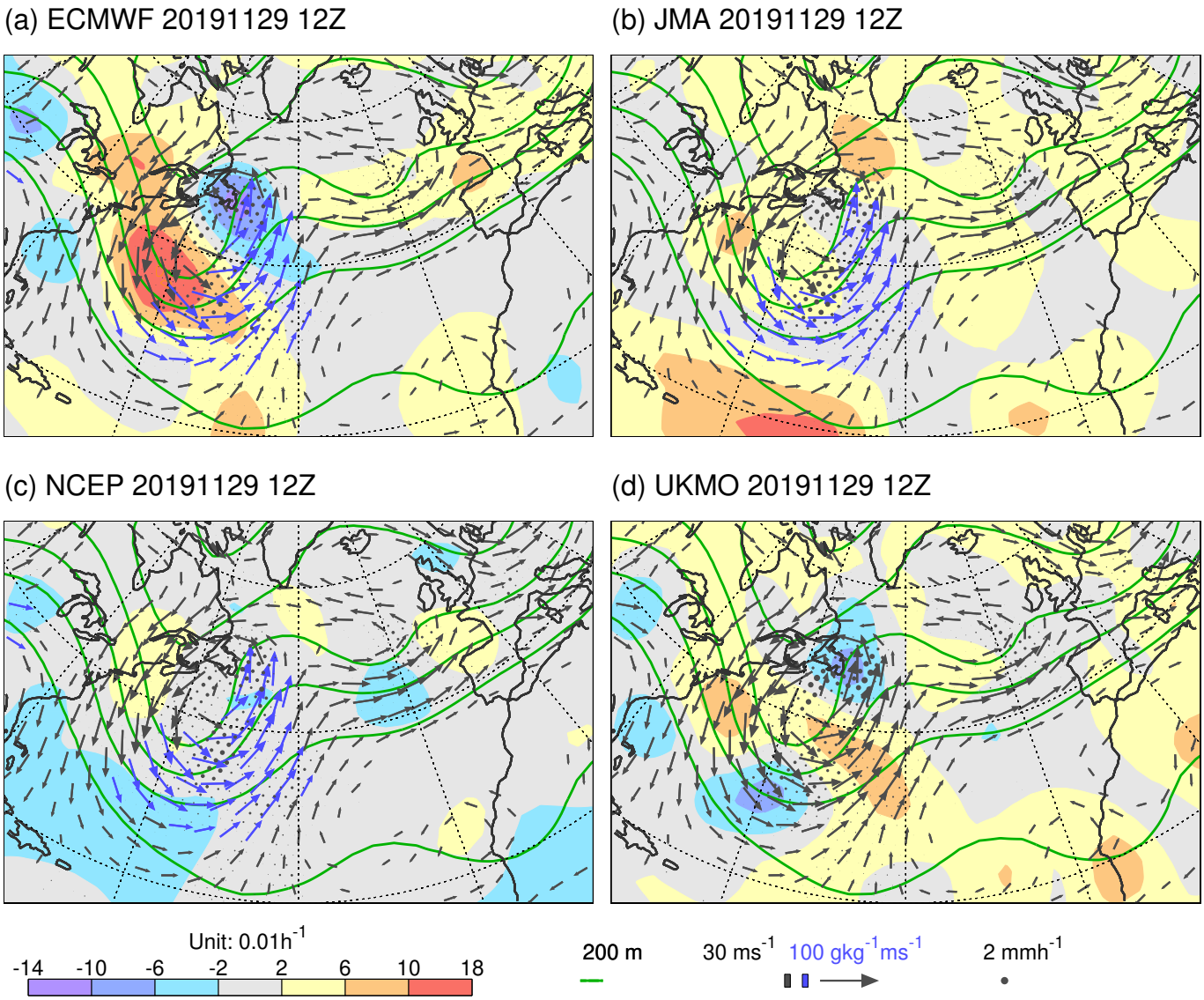

Figure 3. Growth rate LGR_Z for $Z250$ for TIGGE models (shaded), centred at 12 UTC on 29 November 2019. Note that orange contours extend the shading scheme with the same interval, where required. In these cases, the most extreme values are indicated at the ends of the colour bar. Also shown, for the unperturbed ensemble members, are $Z250$ (green contours) and $\mathbf{v}850$ (vectors, which are plotted blue if the moisture flux at 850 hPa exceeds $100\,\mathrm{g\,kg^{-1}m\,s^{-1}}$). Ensemble mean precipitation is indicated with (black dots, with radius proportional to the precipitation rate up to the maximum size at $2\,\mathrm{mm\,h^{-1}}$). The models shown are from (a) ECMWF, (b) JMA, (c) NCEP, and (d) UKMO as discussed in Sect. 2.2. The fields shown are constructed using the first 12 h of each model's ensemble forecasts started at 00 and 12 UTC. Calculations use centred-means and differences between consecutive 6-hourly leadtimes (0 h,6 h,12 h). All other details are the same as in the caption to Fig. 2, except that the plotted $Z250$ field is not spatially smoothed (it is considered already a synoptic scale field). Note that humidity fluxes could not be calculated for the UKMO model as specific humidity data was not available from TIGGE.

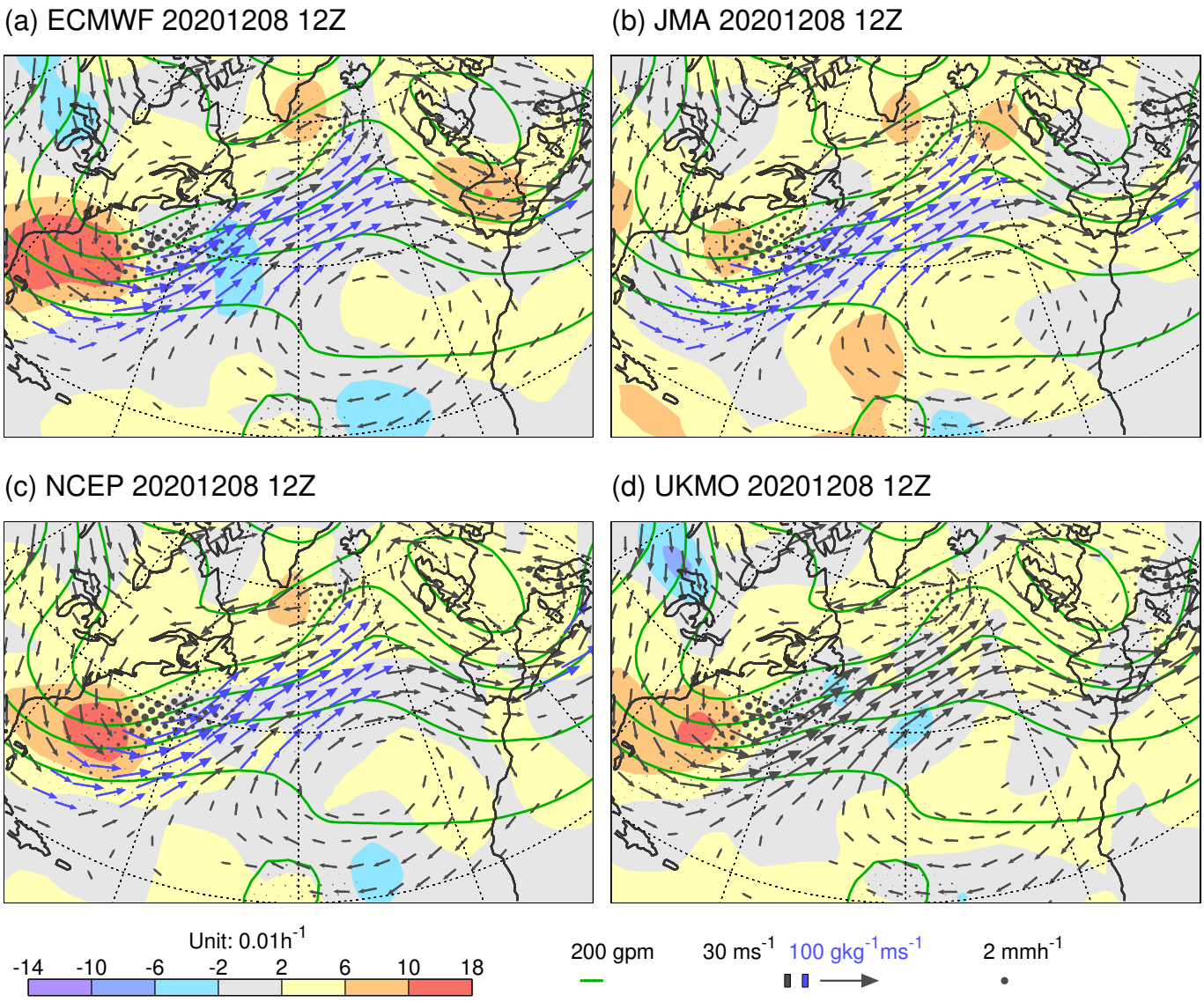

Figure 4. As Fig. 3, but centred at 12 UTC on 8 December 2020. This situation also corresponds to an event of cyclone intensification off the North American east coast associated with a WCB.

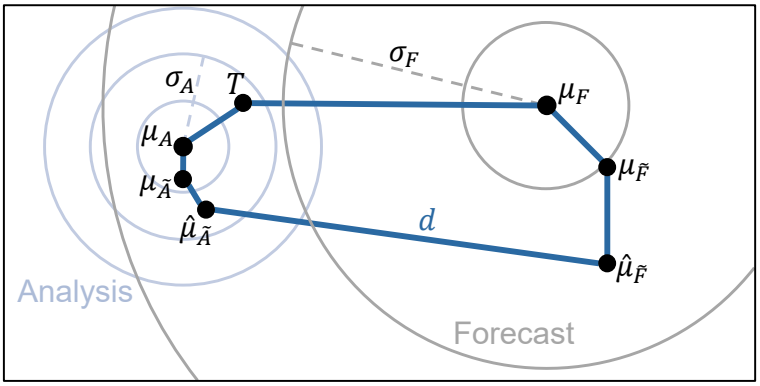

Figure 5. Schematic in 2 dimensions showing the "departure" d as the difference between the ensemble mean of the operational forecast $\widehat{\mu}_{\widetilde{F}}$ and the ensemble mean of the verifying operational analysis $\widehat{\mu}_{\widetilde{A}}$. The forecast and analysis ensembles can be considered as finite samplings of the underlying distributions with (mean,variance) $= (\mu_{\widetilde{F}}, \sigma^2_{\widetilde{F}})$ and $(\mu_{\widetilde{A}}, \sigma^2_{\widetilde{A}})$, respectively. Circles depict the hypothetical forecast and analysis distributions with (mean,variance) $= (\mu_F, \sigma^2_F)$ and (μ_A, σ^2_A), respectively, that would be created with a perfect model (and imperfect/incomplete observational information). The truth is indicated with a T.

For the interested reader, Sect. 4.1 gives a derivation of an extended spread–error equation, which takes bias, analysis uncertainty and sampling into account in the evaluation of ensemble spread. Salient aspects of this extended spread–error equation are discussed in Sect. 4.2.

4.1 Derivation of the extended spread–error equation

For initial time t_j, the departure of the ensemble mean forecast $\widehat{\mu}_{\widetilde{F}j}$ from the ensemble mean analysis $\widehat{\mu}_{\widetilde{A}j}$ can be written (by following the other solid blue lines in Fig. 5) as:

$$\begin{aligned} d_j &= (\widehat{\mu}_{\widetilde{F}j} - \widehat{\mu}_{\widetilde{A}j}) \\ &= (\widehat{\mu}_{\widetilde{F}j} - \mu_{\widetilde{F}j}) + (\mu_{\widetilde{F}j} - \mu_{Fj}) + (\mu_{Fj} - T_j) + (T_j - \mu_{Aj}) + (\mu_{Aj} - \mu_{\widetilde{A}j}) + (\mu_{\widetilde{A}j} - \widehat{\mu}_{\widetilde{A}j}) \\ &= \underbrace{(\mu_{Fj} - T_j)}_{\mathbf{1}} + \underbrace{(\widehat{\mu}_{\widetilde{F}j} - \mu_{\widetilde{F}j})}_{\mathbf{2}} + \underbrace{(T_j - \mu_{Aj})}_{\mathbf{3}} + \underbrace{(\mu_{\widetilde{A}j} - \widehat{\mu}_{\widetilde{A}j})}_{\mathbf{4}} + \underbrace{(\mu_{\widetilde{F}j} - \mu_{Fj})}_{\mathbf{5}} + \underbrace{(\mu_{Aj} - \mu_{\widetilde{A}j})}_{\mathbf{6}} \quad , \end{aligned} \tag{4}$$

where the last line is just a convenient re-arrangement of the terms on the second line.

As discussed above, the terms **1**–**6** in Eq. (4) would be difficult to quantify in the operational forecast system because they require knowledge of the truth and the moments of the underlying distributions (only d_j, $\widehat{\mu}_{\widetilde{F}j}$ and $\widehat{\mu}_{\widetilde{A}j}$ are quantifiable). Nevertheless, we can discuss their expected values. The expectation operator $\mathbb{E}_j[\cdot]$, introduced above, is the expectation for a given initial time t_j. The forecast distributions F_j and $\widetilde{F}_j$ are fixed for given initial time, so the expectation $\mathbb{E}_j$ is over

the potential T_j (with distribution F_j) and the potential analysis distributions A_j and $\widetilde{A}_j$ (which are dependent on T_j). The expectation is also over the finite ensemble samplings $\widehat{F}_j$ of $\widetilde{F}_j$ and $\widehat{A}_j$ of $\widetilde{A}_j$.

The reliability of $A_j(T_j)$ implies that $\mathbb{E}_j[T_j] = \mathbb{E}_j[\mu_{Aj}]\,(= \mu_{Fj})$ and $\mathbb{E}_j[(T_j - \mu_{Aj})^2] = \mathbb{E}_j[\sigma^2_{Aj}]$. Taking the expectation of the last line in Eq. (4) we have:

$$\mathbb{E}_j[d_j] = \mathbb{E}_j\left[\mathbf{5}+\mathbf{6}\right] = \mathbb{E}_j\left[(\mu_{\widetilde{F}j} - \mu_{Fj}) - (\mu_{\widetilde{A}j} - \mu_{Aj})\right] \equiv \beta_{\widetilde{F}j} - \beta_{\widetilde{A}j} \equiv \beta_j \quad , \tag{5}$$

since terms **1**–**4** in Eq. (4) have zero expectation. Hence $\mathbb{E}_j[d_j] = \beta_j$: the expected bias of the unreliable forecast $\beta_{\widetilde{F}j}$ minus the expected bias of the unreliable analysis $\beta_{\widetilde{A}j}$. As lead-time increases, one might expect that the forecast bias will become the dominant component of β_j.

The extended spread–error equation is based on the expected square of Eq. (4). This involves squared terms, such as $\mathbb{E}_j[\mathbf{1}\cdot\mathbf{1}]$, and cross terms, such as $2\mathbb{E}_j[\mathbf{1}\cdot\mathbf{6}]$. With the squared terms presented in the same order as in the last line in Eq. (4) the expected square of Eq. (4) can be written as:

$$\begin{aligned}\mathbb{E}_j[d_j^2] &= \sigma^2_{Fj} + \frac{1}{m}\sigma^2_{\widetilde{F}j} + \sigma^2_{Aj} + \frac{1}{m}\sigma^2_{\widetilde{A}j} + (\beta_{\widetilde{F}j} - \beta_{\widetilde{A}j})^2 + \mathcal{E}_j \\ &= \frac{m+1}{m}(\sigma^2_{\widetilde{F}j} + \sigma^2_{\widetilde{A}j}) + \beta_j^2 + \Big\{\underbrace{(\sigma^2_{Fj} - \sigma^2_{\widetilde{F}j}) + (\sigma^2_{Aj} - \sigma^2_{\widetilde{A}j})}_{\text{Variance deficit}} + \mathcal{E}_j\Big\} \quad .\end{aligned} \tag{6}$$

The term in parentheses { } comprises the "variance deficit" (which compares the reliable and unreliable variances of the ensemble forecast and analysis) and $\mathcal{E}_j$, which collects any potentially non-zero cross terms and the (expected) variance in analysis bias (see later for further discussion).

From Eq. (6), the expected squared departure (over an infinite set of initial forecast times t_j) can then be written as

$$\mathbb{E}[d^2] = \frac{m+1}{m}\mathbb{E}[\sigma^2_{\widetilde{F}} + \sigma^2_{\widetilde{A}}] + \mathbb{E}[\beta]^2 + \mathbb{E}[R] \quad , \tag{7}$$

with expected "residual":

$$\mathbb{E}[R] = \mathbb{E}\left[(\sigma^2_F - \sigma^2_{\widetilde{F}}) + (\sigma^2_A - \sigma^2_{\widetilde{A}})\right] \;+\; \mathbb{E}[\mathcal{E}] \;+\; \sigma^2_\beta \quad , \tag{8}$$

where the variance in forecast bias $\sigma^2_\beta = \mathbb{V}[\beta]$ accounts for the explicit replacement of $\mathbb{E}[\beta^2]$ with $\mathbb{E}[\beta]^2$ in Eq. (7). It can be seen that all terms that involve variations in bias have been moved into the residual.

For an unbiased estimator of Eq. (7), we note that $\left(\mathbb{E}[d^2] - \mathbb{E}[\beta]^2\right) = \left(\mathbb{E}[d^2] - \mathbb{E}[d]^2\right) = \mathbb{V}[d]$ which has unbiased estimator $\frac{n}{n-1}\left(\overline{d^2} - \overline{d}^2\right)$, and obtain the extended spread–error equation:

$$\underset{\text{Error}^2}{\frac{n}{n-1}\overline{d^2}} = \underset{\text{Spread}^2}{\frac{m+1}{m-1}\overline{\left(\widehat{\sigma}^2_{\widetilde{F}} + \widehat{\sigma}^2_{\widetilde{A}}\right)}} + \underset{\text{AnUnc}^2}{\frac{n}{n-1}\overline{d}^2} + \underset{\text{Bias}^2}{\overline{R}} \quad \underset{\text{Residual}^{(2)}}{} \quad , \tag{9}$$

where the various terms have been named for future reference. For the purposes of calculating statistical significance, $\overline{R}$ is written as the mean of the residuals R_j:

$$R_j \equiv \frac{n}{n-1}\left(d_j^2 - \overline{d}^2\right) - \frac{m+1}{m-1}\left(\widehat{\sigma}^2_{\widetilde{F}j} + \widehat{\sigma}^2_{\widetilde{A}j}\right) \quad , \tag{10}$$

which close the budget for each initial time (for given, constant $\overline{d}$). The extent to which $\overline{R}$ is a good estimate for the expected "variance deficit" in Eq. (8) will depend on the magnitude of the other terms $\mathbb{E}\left[\mathcal{E}\right] + \sigma^2_{\beta}$ in Eq. (8). In Appendix A, it is suggested that the only term that could be non-negligible is the variance in forecast bias, σ^2_{β}.

4.2 Summary of the extended spread–error equation

The Error^2 term in Eq. (9) is the scaled mean square of the departures d. (Scaling factors within each term ensure that the equation is valid for any number of forecasts $n > 1$ and any ensemble size $m > 1$). Eq. (9) writes the Error^2 as the sum of the mean estimated forecast variance (Spread^2), mean estimated analysis variance (AnUnc^2), squared estimated mean bias (Bias^2) and the Residual$^{(2)}$. This final term simply closes the budget — the superscript $^{(2)}$ signifies that Residual$^{(2)}$ is also in squared units, although it can be negative as well as positive. For the purpose of statistical significance testing, Residual$^{(2)}$ can be written as a mean of residuals Eq.(10), which close the budget for each initial time (assuming constant bias).

If Bias $\overline{d}$ is different from zero, then this indicates a bias in the mean of the underlying forecast distributions relative to that of the underlying analysis distributions. Residual$^{(2)}$ estimates the sum of the deficits in forecast and analysis variance, plus the (potentially non-negligible) temporal variance in forecast bias. A negative Residual$^{(2)}$ indicates a surplus in variance, while positive values *can* indicate a deficit in variance provided the variance of forecast bias is negligible or accounted for; see later. Hence, if either Bias or Residual$^{(2)}$ differ from zero, then this indicates a lack of reliability in either the first or second moments of the forecast or analysis distributions. As lead-time increases, one might expect the issues in the forecast distribution to begin to dominate.

For display purposes the terms in Eq. (9) can be put into more understandable units by taking their square-roots — signified by removing the superscript ^{2} label. This approach leaves the bias in its correct form, for example. Since $\overline{R}$ can be positive or negative, "Residual" $= \sqrt{|\overline{R}|} \cdot \text{SGN}(\overline{R})$ is plotted to retain the sign. While smaller terms will look more important than they are in the squared budget, the residual still correctly indicates spread deficiencies. Statistical significance is determined (at the 5% level with a t-test) from the un-rooted terms (except for Bias^2, which must be determined in its non-squared form).

As an aside, a very similar equation to Eq. (9) was presented by Rodwell et al. (2016) for application in "observation space". That equation represents a more fundamental "ideal" where assigned observation error variances are assumed to represent their true values. In reality, these observation error variances can be inflated to account for representativeness error (Rennie et al., 2021), observation error correlations, or when associated with non-linear observation operators. This can mean that the budget will not balance for a given observation type, even if the resulting analyses and forecasts are reliable in "model space". Hence the current model space application represents a complementary approach, which can be more readily applied to a range of models and lead-times, and which can be used to assess ensemble initialisation aspects.

4.3 Seasonal mean forecast reliability in the TIGGE ensembles

Figure 6 shows the (square-rooted) terms of the extended spread–error Eq. (9) for $Z250$ based on all day–2 ensemble forecasts verifying in DJF 2020/21 from the ECMWF (top row), JMA (second row), NCEP (third row) and the UK Met Office (bottom row) ensembles. Focusing first on ECMWF (top), the North Atlantic winter stormtrack is evident as a region of enhanced ensemble spread (Fig. 6b). Even without the AnUnc and Bias contributions, the Spread is larger than required to balance the Error (Fig. 6a) — signifying "over-spread" in the stormtrack at day 2. Analysis uncertainty is also enhanced in the stormtrack region (Fig. 6c) and a statistically significant Bias is seen over the east coast of North America. Confirmation of the over-spread is seen with the large negative Residual (Fig. 6e). Note the different colour-bar convention for the Bias and Residual, which can take positive and negative values. Over the North American east coast, the squared budget Eq. (9) is roughly $15^2 = 18^2 + 3^2 + 6^2 - 12^2$ m^2. Because AnUnc and Bias are not negligible here, the 'reliable Spread' (i.e., the Spread required for a zero Residual) in this region would actually be $\sim 13.4^2$ m^2, somewhat less than the 15^2 m^2 which might be inferred from the standard spread–error relationship. Accounting for the variance of forecast bias (which is difficult to estimate here but see later) would further reduce the reliable Spread a little. In contrast, there is a positive Residual over the subtropical North Atlantic (Fig. 6e). Whether this indicates insufficient Spread to account for the elevated Errors in this region (Fig. 6a) or is associated with variance in forecast bias is also discussed later. Note that Rodwell et al. (2018) indicated better reliability for this model. Partly this reflects compensation in their annual and hemispheric means, partly it reflects the importance (here) of accounting for bias and analysis uncertainty, and partly it reflects a recent deterioration in stormtrack reliability.

For the JMA ensemble during this season, day–2 Error is larger than for ECMWF (cf Fig. 6a,f), and the Spread is increased by a larger amount (cf Fig. 6b,g). Note that mean values and root-mean-square (RMS) values, integrated over the area shown, are indicated above each panel in Fig. 6 — it is the RMS values which are most appropriate for comparison. AnUnc and Bias are also larger for JMA (cf Fig. 6c,h and d,i). Consequently, the residual is more strongly negative (cf Fig. 6e,j) — indicating more severe over-spread in this ensemble at this leadtime. For the NCEP ensemble relative to ECMWF, a smaller increase in Spread (cf Fig. 6b,l) than in Error (cf Fig. 6a,k) leads to better variance reliability (cf Fig. 6e,o), despite having larger AnUnc and Bias (cf Fig. 6c,m and d,n). For the UKMO ensemble relative to ECMWF, Error is larger (cf Fig. 6a,p) and Spread is reduced (cf Fig. 6b,q) — leading to the best variance reliability of the four models (Fig. 6t); again despite having larger AnUnc and Bias (cf Fig. 6c,r and d,s) than for ECMWF. Note that conclusions drawn in this section appear to generalise to other parameters (such as geopotential heights and temperatures at 500 hPa), other seasons, other stormtracks, and continue until the

most recent check for the March — May season 2022 (not shown). Using these four models as a demonstration of what could be possible in day 2 ensemble forecasts for the North Atlantic stormtrack, the conclusion would be to pick the errors, analysis uncertainties and bias of the ECMWF model, and the spread and reliability of the UKMO model. With reference to Table 1, an interesting commonality of the two most over-spread systems (ECMWF and JMA) is the use of singular vector perturbations in their initial conditions. Puzzlingly, however, JMA appears to show the weakest initial growth rates.

So, a partial answer to the question in Sect. 3.4 is yes, initial growth rates in the ECMWF ensemble do seem to be too strong within the winter North Atlantic stormtrack. To fully answer the question, it needs to be determined whether the negative Residual in Fig. 6e is associated with a general level of over-spread or whether it can be linked to cyclogenesis events per se? The aim now is to address this question by clustering the stormtrack flow configurations in DJF 2020/21. In the next section, the method of clustering is discussed.

4.4 Compositing cases of cyclogenesis in the ECMWF ensemble

The aim is to obtain a cyclogenesis/non-cyclogenesis partition of the 12 hourly analysed synoptic flow within the DJF 2020/21 season. The method used is *K*-means clustering (Hartigan and Wong, 1979), which seeks to minimise the sum of squared deviations from the relevant cluster-mean. It is used here to obtain three flow clusters in a prescribed region of the North Atlantic stormtrack. The data used within the clustering come from the ECMWF forecasts initialized at 00 and 12 UTC, with $Z250$, and zonal and meridional wind at 850 hPa ($u850$, $v850$) at step 0 from the control forecast during DJF 2020/21, and ensemble mean 12 h accumulated total precipitation from the previous forecast (so the end of the accumulation period corresponds to the time of the circulation fields). It is thought that these fields should be able to capture upper-tropospheric Rossby waves, baroclinic structures and associated diabatic processes. It is the ability to cluster on structures which motivated the choice of the *K*-means approach. Three clusters were thought to provide sufficient degrees of freedom to differentiate the local synoptic-scale structures, while giving large-enough clusters to obtain statistical significance. Each field is on a regular F32 (~ 300 km) Gaussian grid, standardised (about its area- and temporal-mean) and root-cos-latitude-weighted prior to application of the clustering algorithm (total precipitation is standardised by dividing by the square root of its area- and temporal-mean-squared value). This is to give approximately equal weight to each field and sub-region. Three random date/times (out of the 180 date/times available within the season) are used to initialise the clusters. Since there is no guarantee that the algorithm will identify the optimal solution, 100 initialisations were performed and the same minimum sum of squared deviations from the cluster-means was generally found within the first 5 such initialisations — indicating that the solution is optimal.

The first clustering region is located at the head of the North Atlantic stormtrack [80°–50°W, 30°–50°N] and contains 11×7 data points on the F32 grid. The rationale for choosing this region is that it corresponds to the North Atlantic "hot spot" for cyclone intensification (e.g. Wernli and Schwierz, 2006) and WCB activity (Madonna et al., 2014), and its size corresponds to a half-wavelength of a typical baroclinic wave. Figure 7 (top row) shows the same fields as in Fig. 3 but averaged over the three sets of date/times obtained from the *K*-means clustering in this (indicated) region. Cluster 1 (Fig. 7a, 32 date/times) appears to capture a partially-evolved cyclogenesis flow-type off the east coast of North America, with a baroclinic westward tilt with height, intense horizontal moisture flux and precipitation ahead. A general tendency for strong uncertainty growth rate at the

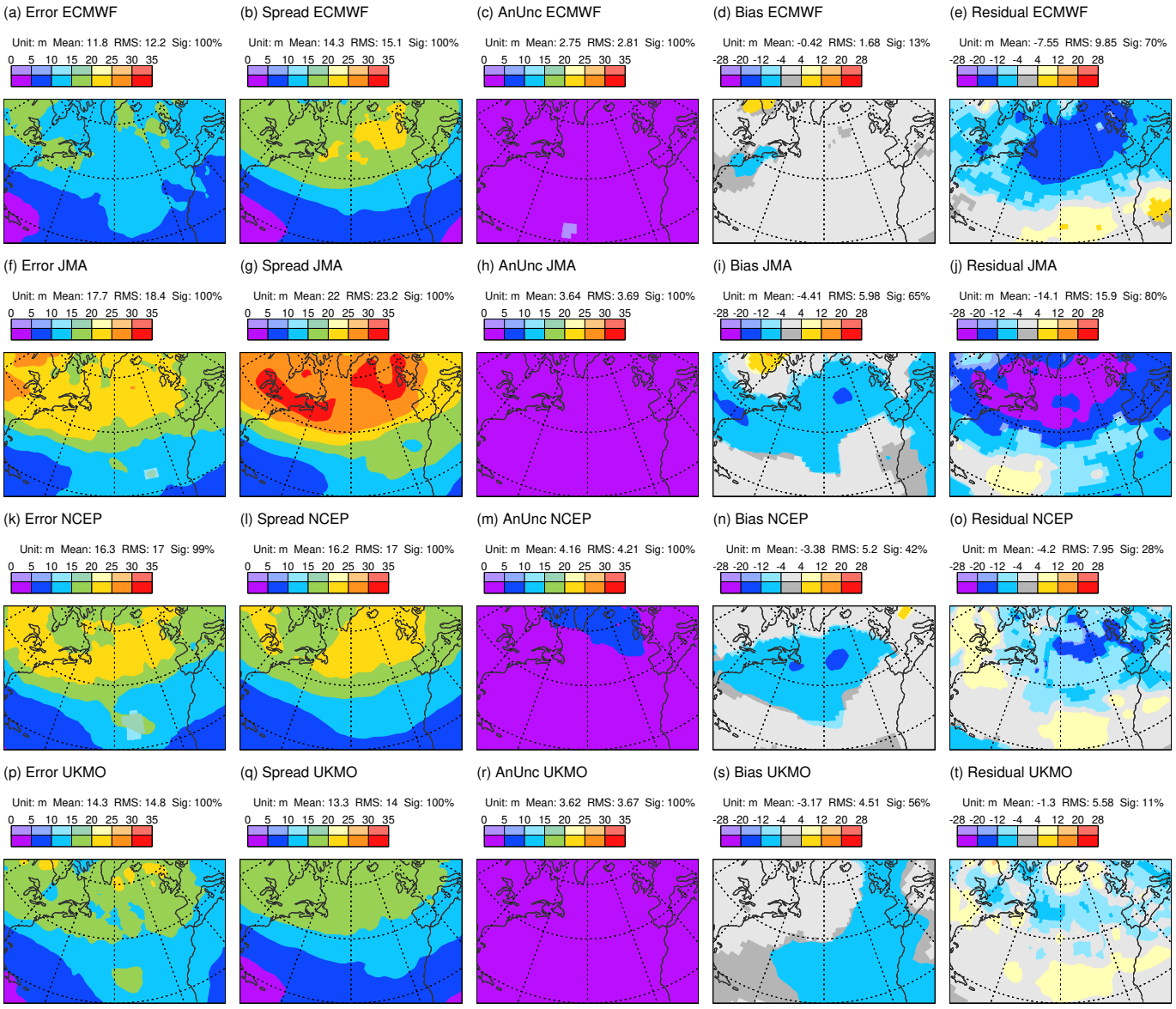

Figure 6. Square-roots of terms in the extended spread–error budget Eq. (9) for $Z250$ based on all day 2 forecasts verifying in DJF 2020/21. Note that Residual $= \sqrt{|R|}\mathrm{SGN}(R)$ to retain the correct sign. Data comes from the TIGGE archive for ECMWF (top row), JMA (second row), NCEP (third row) and UKMO (bottom row) ensembles. Statistically significant values are shown with more saturated colours. Area means (for the area displayed) are indicated at the top of each panel (e.g. "Sig: 85%" means that 85% of the area shown is significant at the 5% significance level, using an auto-regressive AR(1) model to take account of serial correlation).

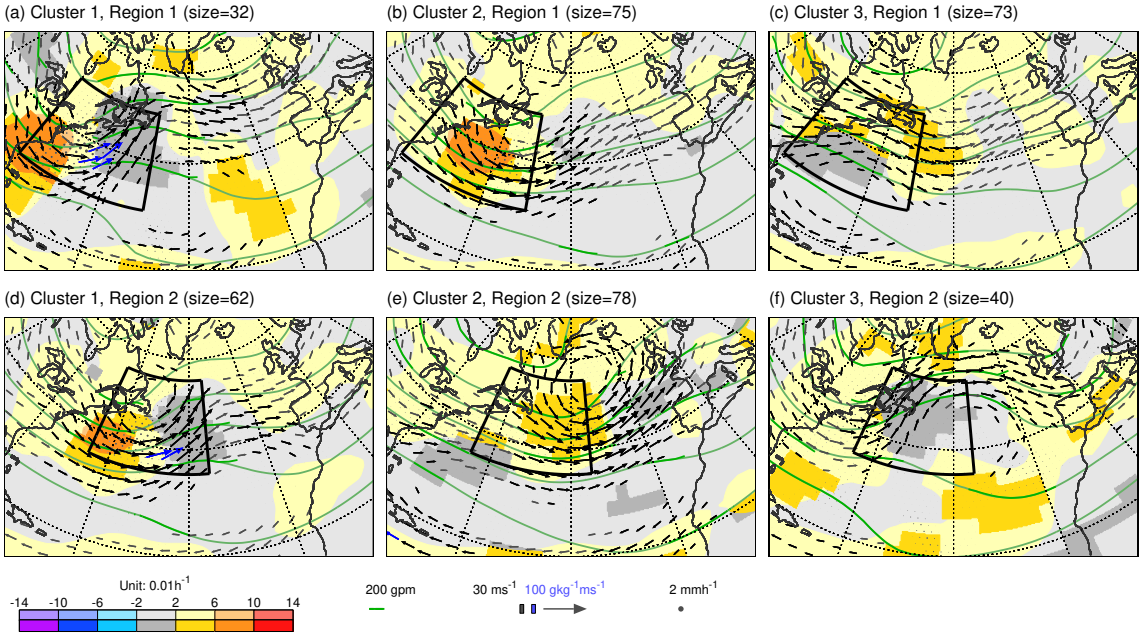

Figure 7. Means over the date/times for the three clusters obtained from K-means clustering in the first (top) and second (bottom) region on the fields of $Z250$ (contours), $u850$ and $v850$ (vectors) from ECMWF control (unperturbed) forecasts at step 0, and ensemble mean 12 h accumulated precipitation (dots) valid (or end of accumulation) at 00 and 12 UTC within the period DJF 2020/21. The two clustering regions are indicated by the black borders in each panel. Although not used in the clustering, shading shows the corresponding mean 12 h uncertainty growth rate LGR_Z for $Z250$. Vectors are coloured blue when the humidity flux at 850 hPa exceeds 100 g kg^{-1} m s^{-1}. More saturated shading, contours, vectors and dots indicate statistical significance at the 5% level (not accounting for autocorrelation due to the discontinuous nature of date/times in each cluster).

southern extent of the upper-level trough is also evident. (Note that the growth rate is not used within the clustering algorithm since this could potentially bias the reliability assessment). Cluster 2 (Fig. 7b, 75 date/times) shows a broader trough, weaker moisture flux, and possible cyclogenesis, displaced further downstream. Cluster 3 (Fig. 7c, 73 date/times) shows a diffuse ridge with a trough even further downstream, and a surface anticyclone in the subtropical western North Atlantic.

To identify cyclogenesis events further downstream in the stormtrack, a second clustering region is displaced north-eastward to [65°–35°W, 35°–55°N]. This region also contains 11×7 data points. Clustering results for this region are shown Fig. 7 (bottom row). Cluster 1 (Fig. 7d, 62 date/times) highlights further cyclogenesis with a closed cluster-mean circulation at 850 hPa over Newfoundland and with strong growth rates. Note that, as might be expected, 44 of these 62 date/times (71%) were in cluster 2 for region 1 (Fig. 7b).

By combining the two cyclogenesis clusters (Fig. 7a and Fig. 7d), a total of 91 date/times were identified as cyclogenesis flow-types (32+62 minus 3 duplicates). For each of these date/times (and their 180-91=89 counterpart date/times), visual inspection of plots similar to those in Fig. 1 suggests that the objective clustering has been successful in partitioning the

date/times into cyclogenesis and non-cyclogenesis flow-types. This then allows the evaluation of the extended spread–error budget for a large set of cyclogenesis events and of the counterpart set.

4.5 Forecast reliability during cyclogenesis in the ECMWF ensemble

Because Fig. 7a and Fig. 7d show partially evolved cyclogenesis flow-types, it is necessary to wind-back the date/times a little to evaluate the day–2 extended spread–error budget during cyclogenesis. Winding back by one and two days gives very similar results (not shown). Also, conclusions are very similar for the evaluation based on the date/times obtained for the first region alone; albeit with the impact more confined to the western end of the stormtrack. Here, results are shown for a 2–day wind-back (consistent with the time for moderate deepening of a low-pressure system; Wernli and Davies, 1997) and for both regions together. This means that the cyclogenesis and counterpart composites represent a 91:89 partition (nearly 50:50) of the data used in Fig. 6 (top row).

Figure 8 shows the (square-roots of) the terms in the extended spread–error equation Eq. (9) for $Z250$ at day 2 in the ECMWF ensemble, separately for cyclogenesis (top) and counterpart (middle) composites, and their difference (bottom). The black border indicates the union of the two clustering regions. Comparison shows that ensemble spread for the cyclogenesis composite (Fig. 8b) is enhanced in the western part of the North Atlantic stormtrack while the spread for the counterpart composite (Fig. 8g) is centred more downstream. There are also corresponding differences in analysis uncertainty (Fig. 8c and Fig. 8h), possibly due to differing uncertainty growth rates in the background forecasts used within the ensemble data assimilation process.

Figure 8n indicates a significant flow-dependent difference in mean absolute forecast bias (of about 6 m) along the eastern coast of North America. This equates to a variance in forecast bias of $\sim 3^2$ m^2 — suggesting that this term is as important in (the Residual term of) the day 2 extended spread–error budget for the whole of DJF 2020/21 as the explicitly represented analysis uncertainty (Fig. 6c), and would suggest an even stronger over-spread issue (cf. term estimates in Section 4.3, with a revised 'reliable spread' of $\sim 13.1^2$ m^2). In contrast, the positive Residual term over the subtropical North Atlantic (Fig. 6e) could reflect variance in forecast bias ($\sim$Fig. 8n) rather than simply indicating ensemble under-spread. It is possible that the flow-dependent variance in forecast bias might have broader consequences — for example in the development of "weak constraint" approaches to data assimilation, which attempt to account for model bias (Laloyaux et al., 2020). Here, in this flow-dependent evaluation, the inter-cluster variability in forecast bias is explicitly represented in Fig. 8d and Fig. 8i. Notice that the bias along the east coast of North America for the counterpart composite (Fig. 8i) appears to account for the increased errors seen in this region (compare Fig. 8f and Fig. 8a).

The overall assessment of ensemble spread is seen in the residual terms (Fig. 8e and Fig. 8j). Here it is evident for the ECMWF ensemble that most of the over-spread in the region of focus, at the western end of the North Atlantic winter storm-track (Fig. 6e), is associated with the cyclogenesis composite — with statistically significant residuals in Fig. 8e and statistically insignificant residuals (indicated by the light blue and light grey colours) in Fig. 8j. Differences are shown in Fig. 8o. They are particularly strong and significant over Newfoundland. Downstream, differences have the opposite sign — possibly associated with differences in downstream cyclogenesis, and consistent with the increased spread noted above. Linking the day 2

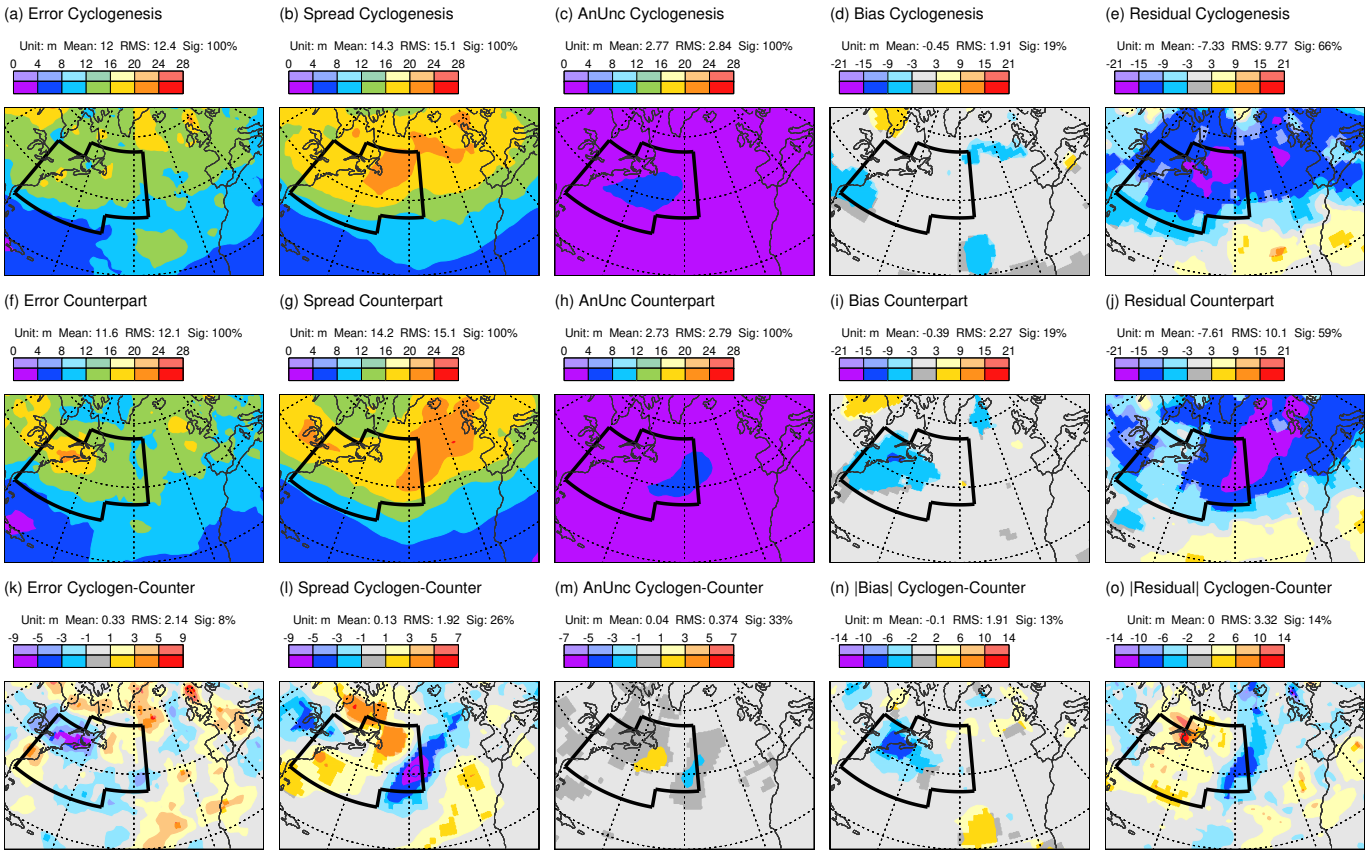

Figure 8. As Fig. 6 but separately for the cyclogenesis cluster date/times (top) and the non-cyclogenesis counterpart cluster date/times (middle), for forecasts that start two days prior to the cluster classifications. The bottom row shows cyclogenesis minus counterpart. Note that differences in absolute Bias and Residual are shown, so that blue (red) colours anywhere on the bottom row show where results for the cyclogenesis composite are better (worse) than for the counterpart. The two clustering regions are indicated by black borders.

stormtrack over-spread (in the region of focus) to cyclogenesis is a key conclusion of this study. It does appear, therefore, that ECMWF initial growth rates (Fig. 2, Fig. 3a) associated with cyclogenesis events are too strong. The next section explores the root-causes for this problem.

5 Sensitivity experiments to quantify sources of uncertainty in the ECMWF ensemble

In Sect. 3, the initial growth rate of uncertainty (Eq. 3) was discussed. It has subsequently been demonstrated that these growth rates are likely to be too strong in the ECMWF ensemble during cases of extratropical cyclogenesis. Here, sensitivity experiments are used to investigate why this growth rate is too strong in the ECMWF ensemble during cyclogenesis — leading to over-spread at day 2.

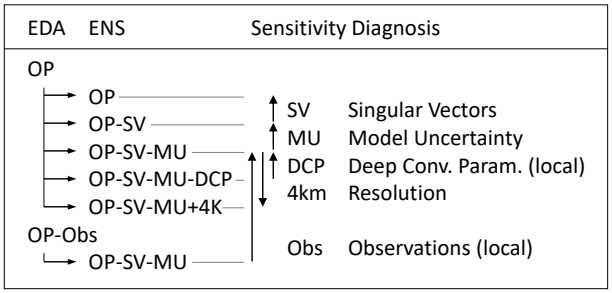

Figure 9. Configuration of IFS sensitivity experiments. These involve the Ensemble of Data Assimilations (EDA) and the Ensemble forecast (ENS). Differences between these configurations allow the diagnosis of sensitivity to individual aspects as indicated. See main text for further details.

Salient details of the ECMWF forecast system were presented in Sect. 2.1. Figure 9 shows the configuration of the sensitivity experiments. It refers to the base "operational" configuration as EDA=OP, ENS=OP. In the sensitivity experiments, this configuration is successively modified. By turning off singular vector perturbations globally (OP-SV), model uncertainty globally (OP-SV-MU) and the parametrization of deep convection in a local box (OP-SV-MU-DCP), or increasing model horizontal grid resolution to ~ 4 km (OP-SV-MU+4km), or not assimilating observations within a given region in the EDA (OP-Obs) and again running the ENS configuration OP-SV-MU. Differences between these configurations allow the diagnosis of individual aspects. Vertical arrows in Fig. 9 indicate the sign convention of the difference to be plotted. The conclusions are not thought to be sensitive to the ordering of the various modifications. For example, it will be seen that the impacts on *total* precipitation of DCP and +4km are small, and hence these impacts should be little changed in the presence of the SPPT form of MU. However, parametrized turbulent fluxes might be weakened with +4km, and hence this impact could be a somewhat different in the presence of MU.

Figure 10 shows day 2 results for the standard deviation (spread) in $Z250$ from sensitivity experiments for the cyclogenesis case initialised at 12 UTC on 28 November 2019 (i.e., the validity time is one day later than that where the growth rate fields were centred in Fig. 2 and Fig. 3a in order to see the combined effect of the uncertainty source and the growth rate). Grey contours show PMSL, and the black contour shows where $P315 = 2$ PVU — highlighting the location of the tropopause on the 315 K isentrope) in the unperturbed EDA analysis. A parallel set of results for spread in $P315$ is shown in Fig. 11. Note that shading intervals vary over the panels shown in these two figures, so that the structures of all impacts can be seen. Mean values and RMS values, integrated over the area shown, are indicated above each panel.

Figure 10a shows the OP configuration with a well-developed surface low pressure system, as discussed in relation to Fig. 1. The WCB associated with this cyclone is seen to lead to the development of a prominent downstream upper-level ridge and a downstream trough west of Europe. As might be expected, the maximum $Z250$ spread is located downstream of the maximum Lagrangian growth rates (cf. Fig. 3a).

The impact of the initial SV perturbations on $Z250$ spread (Fig. 10b) is particularly pronounced along the western flanks of the two prominent troughs over the western and eastern North Atlantic, respectively. This likely indicates the potential for

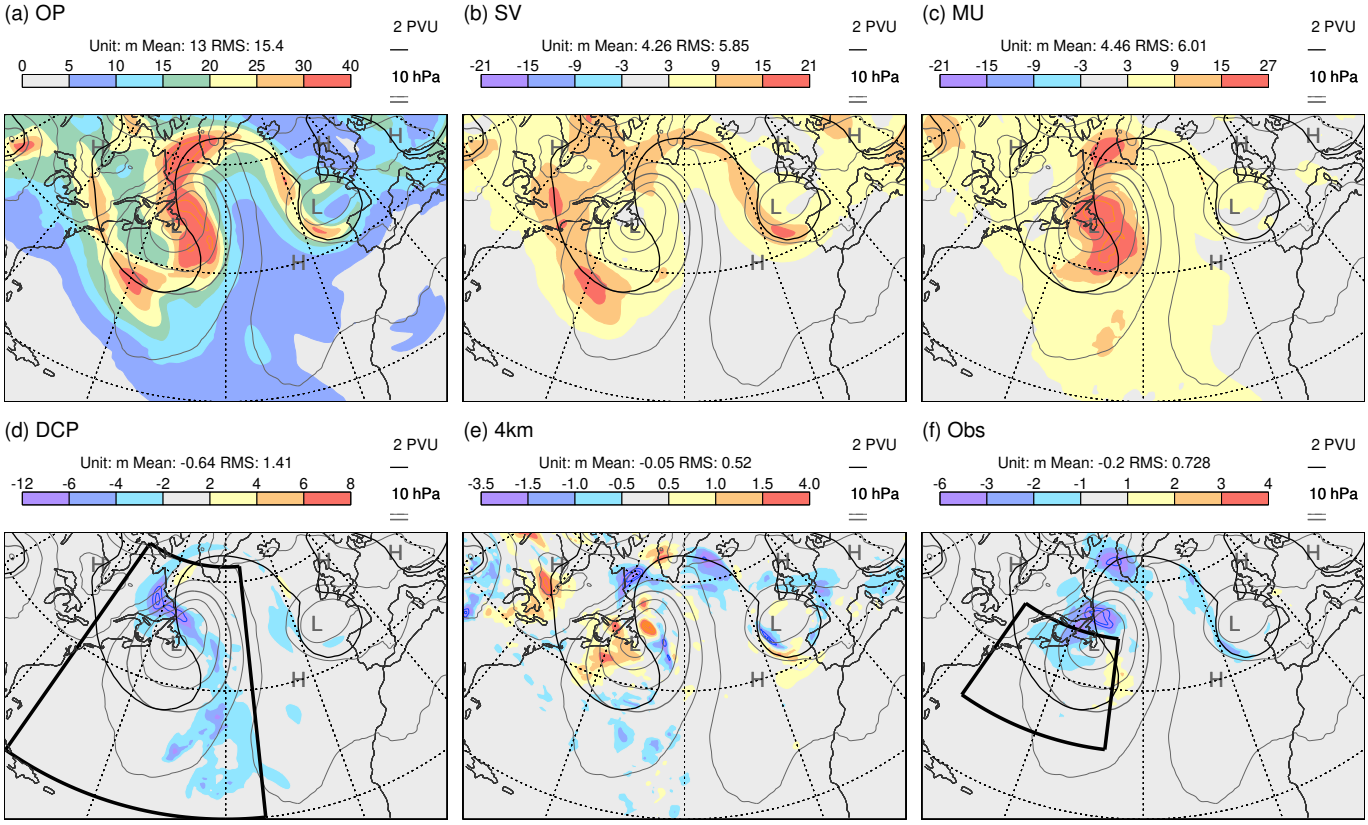

Figure 10. Sensitivity results showing day–2 $Z250$ spread (shaded) from ECMWF ensemble forecast experiments initialised at 12 UTC on 28 November 2019. (a) Total spread for the near-operational configuration OP. (b)–(f) Differences in spread between experiments which highlight the impacts of (b) including initial Singular Vector perturbations, (c) including the Model Uncertainty representation, (d) including the parametrization of Deep Convection in the indicated region [75°W–34°W, 20°N–63°N], (e) an increase in model grid resolution to ~ 4 km, and (f) the assimilation of observations in the indicated region [75°W–47°W, 30°N–49°N]. Contours extend the shading scheme, with the same interval. In these cases, the most extreme values are indicated at the ends of the colour bar. Also shown in each panel are the PMSL (grey contours) and PV=2 PVU on the 315 K isentrope (black contour) from the unperturbed EDA analysis.

dynamic growth along the intense jets in these regions, qualitatively in line with the idealized studies by Hakim (2000). This SV impact might help explain the apparent slight westward shift of the centre of maximum growth in the ENS (Fig. 3a) relative to the EDA (Fig. 2). There are places where the SV impact on spread is half the total (so that the fraction of variance explained reaches 25%). In contrast to the SV impact, the impact of the model uncertainty (MU) representation (Fig. 10c) is particularly pronounced in the cyclone centre and in the region of the WCB ahead of the surface low, i.e., in regions where cloud-related physical processes are particularly active. The large signal along the western flank of the ridge southwest of Greenland is consistent with the results of Joos and Forbes (2016), who found a large influence of cloud microphysical processes in the WCB on the tropopause structure in this part of the downstream ridge. MU also explains up to 25% of the total variance. The remaining variance must be associated with the (deterministic) growth of initial EDA analysis uncertainty.

Fig. 10d shows the impact of including the deep convection parametrization (DCP) in the indicated region (note the smaller contour interval). There is a reduction in the spread — particularly in the WCB region — that would otherwise be created when the model is forced to represent this convection on its 16 km grid. This reduction in spread may go some way to explaining why Rodwell et al. (2018) identified under-spread associated with mesoscale convective systems over North America. Interestingly ensemble mean total precipitation (parametrized plus resolved) is little changed when turning off parametrized deep convection, both in location and amount (not shown).

The impact of increasing the model grid resolution to ~ 4 km is mixed. For $Z250$ spread (Fig. 10e; note the smaller contour interval), the impact is generally weak. In contrast the impact on $P315$ spread (Fig. 11e) is strong, particularly within the WCB region. At 4 km, the model attempts to resolve more of the convection. The resolved convection can be associated with stronger updrafts, which might perturb the tropopause more vigorously, where PV gradients are particularly strong. Given that model uncertainty (MU) representation is thought to partly account for the impact of sub-grid-scale uncertainty, perhaps the most interesting aspect here is the lack of agreement with the MU impact (cf. Fig. 11e and Fig. 11c). The impact on $P315$ uncertainty of allowing the model to resolve more of the convection at the 4 km resolution (Fig. 11e) appears to be in closer agreement with the response to turning off the deep convection parametrisation (minus Fig. 11d), when the model is forced to represent the convection on the 16 km grid. The increase in resolution also results in a small shift of ensemble mean precipitation from parametrized to resolved, with little change in the total (not shown).

The impact of assimilating local observations is obtained using an EDA experiment (OP-Obs, see Fig. 9) where all observations were denied to the EDA in the region [75°W–47°W, 30°N–49°N] for the single data assimilation cycle that generated the initial conditions; observations outside this region are used in both EDA=OP and EDA=OP-Obs. It is not so easy to anticipate the impact of local observations, particularly with 4DVar, when remote observational information can propagate into the region. Results demonstrate that the assimilation of observations in this baroclinic region does reduce initial uncertainty (not shown). Figure 10f and Fig. 11f show the impact at 2 d (note the smaller contour intervals compared to those for the SV and MU impacts). There is a reduction in spread in the developing low, which suggests that the assimilation of local observations (such as cloud-affected radiances) is beneficial in this case. It is less clear whether this impact is simply a sharpening of the distribution, or whether it would also affect growth rates and reliability.

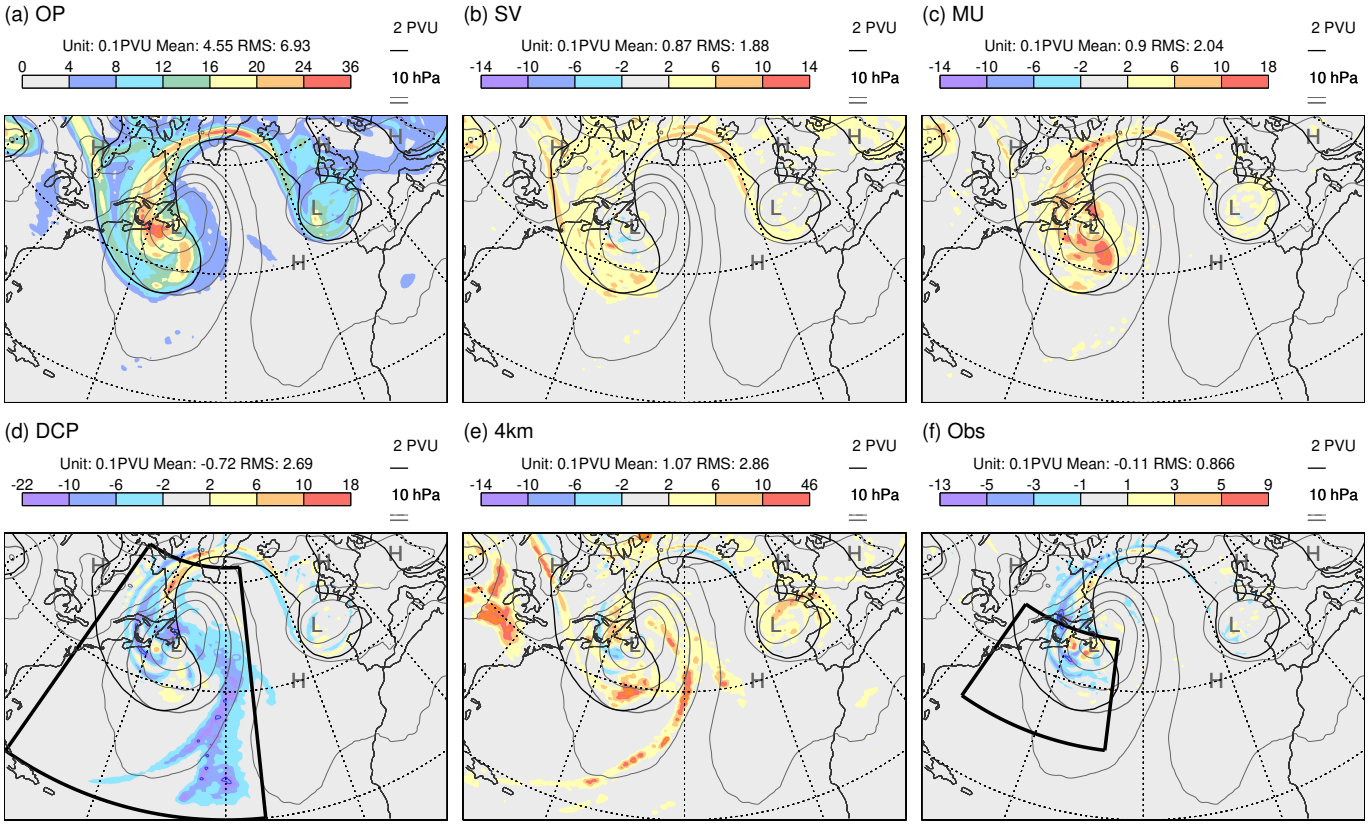

Figure 11. As Fig. 10, but shading shows the spread in $P315$.

Very similar results to those above were obtained for a second set of experiments initialised at 00 UTC on 17 January 2020 — indicating that the conclusions drawn in this section are robust even with only two cases. In particular, the contributions to the total variance from SVs and MU, with the SV impact largely along the western flanks of the trough and the MU impact in the region of the WCB. Again, the parametrization of deep convection acts to reduce spread in the WCB. For $P315$, the impact of increased resolution is again more similar to the effect of turning-off the deep convective parametrization than to that of MU. The main difference of note is that, while there is a reduction in initial spread from the assimilation of local observations, it is weaker than in the case shown above, and its impact at day 2 is marginal.

The above results suggest that SVs, MU, and the deterministic model itself all play important roles in the early development of operational ensemble spread during cyclogenesis. (This is true globally, as can be seen in the variance spectra in Fig. B1 in Appendix B). Since the motivation for the use of SV perturbations (Magnusson et al., 2009), and the initial reason for the development of MU representations (Buizza et al., 1999), was to increase ensemble spread, it makes sense to investigate how, jointly, these aspects might be developed to reduce the over-spread.

6 Conclusions and Discussion

Studies have highlighted a range of flow features over the North American / North Atlantic / European region which can lead to a reduction in weather forecast performance (Rodwell et al., 2013; Riemer and Jones, 2014; Grams and Blumer, 2015; Lillo and Parsons, 2017; Grams et al., 2018; Baumgart and Riemer, 2019) at synoptic scales (Tribbia and Baumhefner, 2004). The Lagrangian growth rate (LGR_P, Rodwell et al., 2018) of potential vorticity, which is routinely used at ECMWF, suggests that initial upper-tropospheric uncertainty growth tends to be orchestrated around such flow features. For example, as part of extratropical cyclogenesis events, the Lagrangian growth rate in the 12 h background forecasts of the Ensemble of Data Assimilations (EDA) highlights strong uncertainty growth at the southern extent of upper-tropospheric troughs, and weaker negative growth rates (ensemble convergence) in the downstream ridge-building region (Fig. 2). This flow-dependent sensitivity to initial uncertainty in the operational forecast is the motivation for the term 'The Cyclogenesis Butterfly' in the title of this study.

The non-linear covariance terms in the generation of LGR_P (Eq. 3) are consistent with the fact that the growth rate will be sensitive to the scales of (initial) uncertainty. Since all scales contribute to EDA variance, it is unlikely that initial forecast growth rates equate directly to the intrinsic growth rates associated with the real atmosphere's sensitivity to small scale perturbations (Durran and Gingrich, 2014). Initial growth rates associated with the 'Cyclogenesis Butterfly' might, however, provide support for the hypothesis of Palmer et al. (2014), that scale interactions and diabatic processes may be partially confined to intermittent synoptic flow types — which could explain the longer than expected intrinsic predictability limit.

Following Baumgart et al. (2019) but with a focus on operational rather than intrinsic predictability, it is interesting to speculate about the processes which could give rise to strong initial growth rates LGR_P in the upper troposphere. These could include uncertainties in, for example, mid-tropospheric latent heating (e.g. Rodwell et al., 2013), deep tropospheric interactions associated with baroclinic (Hoskins et al., 1985) and associated diabatic (Ahmadi-Givi et al., 2004) processes, upper tropospheric divergence associated with dry balanced dynamics (as represented by the "Omega equation" and "Q-vectors"; Sanders and Hoskins, 1990), nonlinear upper-tropospheric dynamics (Baumgart and Riemer, 2019), and local non-conservative processes including near-tropopause radiative forcing (Chagnon et al., 2013). Identification of the most important processes could help in the prioritisation of observational, data assimilation and modelling research. A useful goal, for example, could be to improve the analysis of the fields associated with the strongest synoptic scale initial growth rates. Here, it has been shown that the assimilation of local observational information can sometimes be useful (Fig. 10f, Fig. 11f).

While the EDA displays consistent flow-dependence in initial growth rates, there are differences in 12 h growth rates between ensemble systems in the TIGGE archive (Fig. 3), indicating strong sensitivity to initialisation and modelling aspects (Table 1). ECMWF generally displaying the strongest initial growth rates, and the question arises as to whether this might be 'too strong'. In an attempt to answer this question, the reliability of each ensemble system was evaluated over the December–February 2020/21 season by assessing the consistency between 2 d forecast error and spread (initially for upper-tropospheric 250 hPa geopotential heights) taking into account bias and uncertainty in the verifying analyses. Although the ECMWF ensemble displayed the smallest error, analysis uncertainty and bias (Fig. 6 columns 1,3,4) in the North Atlantic stormtrack, it was

strongly over-spread (Fig. 6e). The UKMO ensemble showed the best reliability (Fig. 6 column 5). These conclusions on reliability appear to generalise to other parameters (such as geopotential heights and temperatures at 500 hPa), other seasons, other stormtracks, and continue until the most recent check for the March — May season 2022 (not shown).

Clustering on flow-types in the western part of the North Atlantic stormtrack (Fig. 7) demonstrated that the ECMWF over-spread in this region was associated with cyclogenesis events (Fig. 8). One consequence is that calibration (possibly based on machine-learning) of ECMWF ensemble spread, without considering whether a cyclogenesis event occurred (or was likely to occur) earlier in the forecast, will likely be a blunt instrument with which to enforce reliability. In contrast, if the root cause of the over-spread can be removed, then the ensemble could gain from improved reliability *and* sharpness (the two key attributes of an ensemble system, Gneiting and Raftery, 2007).

Sensitivity studies with the ECMWF ensemble reveal the sources of uncertainty during cyclogenesis (Fig. 10 and Fig. 11). A large part is associated with the chaotic growth of initial uncertainty within the deterministic model. Results have shown that this growth rate is sensitive to the representation of deep convection in the model. However, singular vector (SV) perturbations to the initial conditions and the model uncertainty (MU) representation are also important. The dry SV perturbations applied at ECMWF, which are optimised for 2–day growth and which target baroclinic instabilities (Molteni and Palmer, 1993), may well be implicated in the over-spread. If SVs are responsible, then this would suggest the desirability to reduce their magnitude when other factors permit. Such a reduction would make the ensemble forecast more consistent over lead-times and permit a better use of initial growth rates in the evaluation and improvement of the model and model-uncertainty — something that would be beneficial throughout the forecast range. Furthermore, sensitivities to switching-off the parametrization of deep convection and to increasing model resolution suggest that the model uncertainty representation should be more strongly focused on convective instabilities (e.g., Christensen et al., 2017). It is possible that such a focus on instabilities (rather than the effects of already-triggered instabilities) might be better explored within the future 'stochastically perturbed parameter' (SPP) framework for model uncertainty — perturbing triggering thresholds for example.

Code and data availability. The key conclusions in this study are derived from data in the TIGGE archive, which is freely accessible. The ERA5 reanalysis data are also freely available. Other data and diagnostic code are available from the authors upon request.

Video supplement. Growth rate animations are available as supplementary material.

Appendix A: Estimating the other terms in the residual

It is useful to write $b_j = \mu_{\widetilde{A}j} - \mu_{Aj}$ (= minus term **6** in Eq. (4)) for the bias in the verifying analysis distribution associated with a given realisation of the truth, and subsequent assimilation of observations. Then, from Eq. (5), the expected bias in the verifying analysis is given by $\mathbb{E}_j[b_j] = \beta_{\widetilde{A}_j}$.

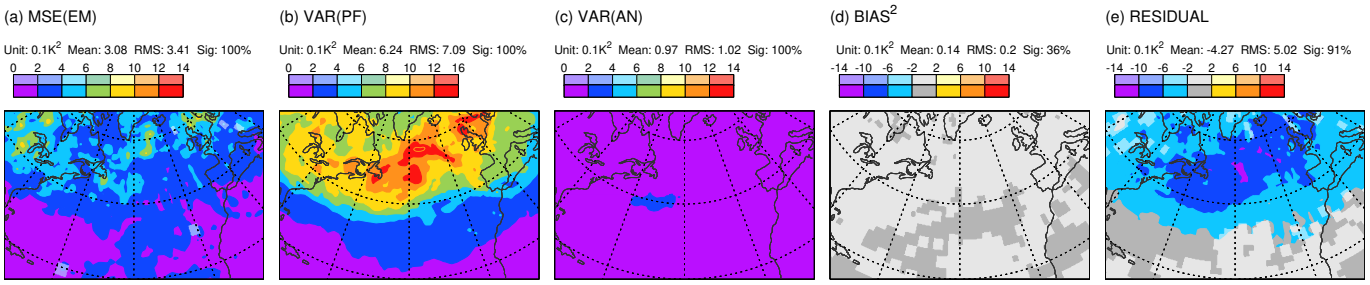

Figure A1. Extended spread–error budget (squared terms) for DJF 2020/21 for temperature at 500 hPa on day 2 from ECMWF ensemble forecasts. Statistically significant values are shown with more saturated colours. Area means (for the area displayed) are indicated at the top of each panel (e.g. "Sig: 85%" means that 85% of the area is significant at the 5% significance level).

With reference to the numbered terms in Eq. (4), one potential cross term contained in $\mathcal{E}_j$ of Eq. (6) is $2\mathbb{E}_j[\mathbf{1}\cdot\mathbf{6}] = 2\sigma_{T_j b_j}$. This might be non-zero if the analysis bias is dependent on the true state (the covariance $\sigma_{T_j b_j}$, when divided by $\sigma^2_{T_j} = \sigma^2_{F_j}$, measures the linear dependence of the analysis bias on the truth). In the ECMWF EDA, each analysis member is an innovation of that member's background forecast, and this lack of independence might imply that the cross-term $2\mathbb{E}_j[\mathbf{2}\cdot\mathbf{4}] = -\frac{2}{m}\sigma_{\widetilde{F}j\widetilde{A}j}$ is non-zero, particularly at very short leadtimes and for small ensemble size m. All other cross-terms are likely to have zero expectation with the exception of $2\mathbb{E}_j[\mathbf{5}\cdot\mathbf{6}] = -2\beta_{\widetilde{F}j}\beta_{\widetilde{A}j}$, which is explicitly represented in the β_j^2 term in Eq. (6). Finally note that $\mathbb{E}_j[\mathbf{6}^2] = \mathbb{V}_j[\mathbf{6}] + \beta^2_{\widetilde{A}j}$ and so $\mathcal{E}_j$ should also include $\mathbb{V}_j[\mathbf{6}] \equiv \sigma^2_{b_j}$, the (expected) variance of the analysis bias. Hence

$$\mathcal{E}_j \approx 2\sigma_{T_j b_j} - \frac{2}{m}\sigma_{\widetilde{F}j\widetilde{A}j} + \sigma^2_{b_j} \quad . \tag{A 1}$$

The expected residual in Eq. (8) can then be written as

$$\mathbb{E}[R] = \mathbb{E}\left[(\sigma^2_F - \sigma^2_{\widetilde{F}}) + (\sigma^2_A - \sigma^2_{\widetilde{A}})\right] \;+\; \mathbb{E}\left[2\sigma_{Tb} - \frac{2}{m}\sigma_{\widetilde{F}\widetilde{A}} + \sigma^2_b\right] \;+\; \sigma^2_\beta \quad , \tag{A 2}$$

As noted in the main text, all terms that involve variations in bias have been moved into the residual. These terms are difficult to estimate, but an attempt is made here for day-2 forecasts of temperatures at 500 hPa during the December–February (DJF) season. For reference, Fig. A1 shows the corresponding (non-rooted version of) the extended spread–error equation. Consistent with the Z500 result in the main text, this also highlights a potential over-spread in the stormtrack. The following list outlines the approach taken here to estimate the right-most four terms in Eq. (A 2).

1. $2\mathbb{E}[\sigma_{Tb}]$ in Eq. (A 2) is tricky to estimate as it requires the state-dependent estimation of analysis bias. One approach is to utilise inter-annual variability over the n_y (=10) years 2012–2021, and to regress the DJF-mean analysis bias (relative to a set of observations) against the DJF-mean observations themselves. The so-called 'analysis departure' is the observation minus the analysis-equivalent of the observation (after applying the relevant 'observation operator'). Assuming that the analysis bias (here for temperatures at 500 hPa, T500) closely matches minus the mean analysis departure from

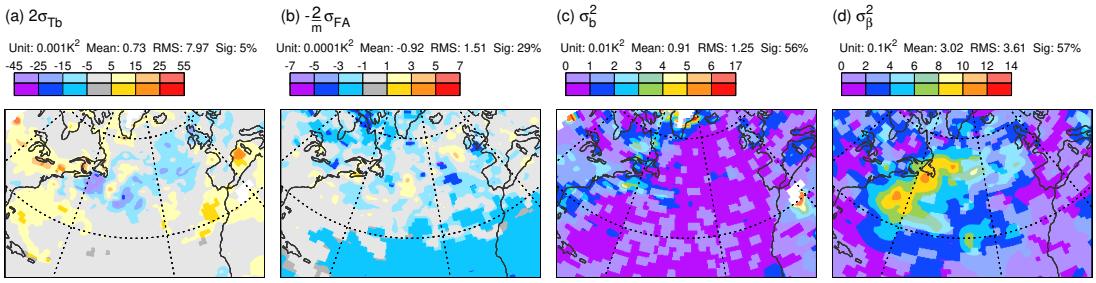

Figure A2. Estimates of the magnitudes of the 'other terms' included in the Residual for temperatures at 500 hPa during the period DJF 2020/21. (a) Covariance of the analysis bias with the atmospheric state ($\times 2$). (b) Covariance of the ensemble-mean day+2 forecast and ensemble-mean analysis ($\times -2$). (c) Variance of the analysis bias. (d) Variance of the day+2 forecast bias (relative to analysis). Statistically significant values are shown with more saturated colours. Area means (for the area displayed) are indicated at the top of each panel (e.g. "Sig: 29%" means that 29% of the area is significant at the 5% significance level). Please see main text for further details.

the observations (here "AMSUA" channel 5 satellite brightness temperatures which measure "mid-tropospheric" temperatures with a maximum weighting at $\sim 500\,\mathrm{hPa}$ - these observations have had 'variational bias correction', VarBC, applied), we can make the estimate

$$2\mathbb{E}\left[\sigma_{Tb}\right] \approx -\frac{2m}{m-1}\frac{\widehat{\sigma}_{o\,(o-a)}}{\widehat{\sigma}_o^2}\overline{\widehat{\sigma}_F^2} \quad , \tag{A 3}$$

where o refers to the seasonal-means of the observations, and a refers to the seasonal-means of the analyses. Notice that the bias variations are scaled to reflect the synoptic variations in truth during DJF 2021. Figure A2(a) shows that, over the North Atlantic, this estimated covariance is $\sim 0.025\,\mathrm{K}^2$. This is generally smaller in magnitude than the analysis variance shown in Fig. A1 (top, third from left, $\sim 0.2\,\mathrm{K}^2$) and more clearly smaller than the current residual (top, far right, $\sim 0.6\,\mathrm{K}^2$). Although the brightness observations have themselves had a bias correction applied, there is a potential that this is not perfect, and that remaining variations in observational bias might affect our analysis bias estimation here. 'Radio-occultation' observations, which are beginning to become more numerous, may offer the prospect of a more pure estimate of analysis bias in the future.

2. σ_b^2, in Eq. (A 2) can also be estimated by utilising inter-annual variability of seasonal-means:

$$\mathbb{E}\left[\sigma_b^2\right] \approx \frac{n'\,n_y}{n_y-1}\widehat{\sigma}^2_{(o-a)} \quad , \tag{A 4}$$

where the multiplier n' is the number of synoptic degrees of freedom in a season (here $n' = 30$ assuming a 3-day synoptic decorrelation timescale). Figure A2(c) shows that this term is also of order $\sim 0.02\,\mathrm{K}^2$ in the North Atlantic region, so again smaller than the analysis variance and residual terms.

3. σ^2_β, in Eq. (A 2) can be estimated by trying a similar approach:

$$\mathbb{E}\left[\sigma^2_\beta\right] \approx \frac{n' n_y}{n_y - 1} \widehat{\sigma}^2_{(f-a)} \quad , \tag{A 5}$$

where f refers to the day+2 seasonal-means of the (deterministic HRES) forecasts. Figure A2(d) shows that this term is of order $\sim 1\,\mathrm{K}^2$ over the western North Atlantic region. This may be an over-estimate since reducing predictive skill with lead-time means that $(f-a)$ begins to reflect (minus) the observed anomaly from climate, and thus becomes less useful at indicating forecast bias as the forecast lead-time increases (e.g. by day+10). Nevertheless, variations in forecast bias may well be non-negligible, and would imply here that the over-spread is even larger over the western North Atlantic. Flow-dependent evaluation of the extended spread–error budget, discussed later, represents a means of avoiding issues with forecast bias variations.

4. $-\frac{2}{m}\sigma_{\widetilde{F}\widetilde{A}}$ in Eq. (A 2) does not involve bias and can be estimated in-sample as

$$\mathbb{E}\left[-\frac{2}{m}\sigma_{\widetilde{F}\widetilde{A}}\right] \approx -\frac{2}{m-1}\overline{\widehat{\sigma}_{\widehat{F}\widehat{A}}} \quad . \tag{A 6}$$

The scale to the colour bar in Fig. A2(b) shows that this term is several orders of magnitude smaller than the analysis variance and residual terms.

Considering all these terms, it appears that the residual in the day+2 T500 budget largely represents the ensemble variance deficit and, possibly, the variance in forecast bias σ^2_β - which is appreciable in the cyclogenesis region off the east coast of North America. Ideally such an analysis would be made for all variables for which the budget is calculated. However, this is difficult for many variables and here we will assume that the T500 analysis carries-over to other variables of the large-scale flow.

Appendix B: Variance spectra for the sensitivity experiments

Figure B1 shows the variance spectra of global spherical harmonics against total wavenumber for the experiments discussed in Sect. 5. The approximate scale is indicated on the top x-axis. Initial uncertainty from the EDA (Fig. B1, thin green curve) occurs at all scales, with the largest variance contributions at scales $\sim 400\,\mathrm{km}$. The thin black curve shows the initial spectrum after addition of the singular vector (SV) perturbations (at scales $\geqslant 1000\,\mathrm{km}$). The contributions to day 2 variance associated with SV perturbations and model uncertainty (MU) representation are indicated by the red and green shaded regions, respectively. SVs contribute strongly at synoptic scales while MU contributes at synoptic and planetary scales. The thick green curve indicates the impact of pure chaotic growth of EDA uncertainty without SVs or MU in the forecast). Notice that the day 2 and EDA variance spectra conincide at scales smaller than $\sim 100\,\mathrm{km}$ — indicating that the EDA variance is already saturated at these

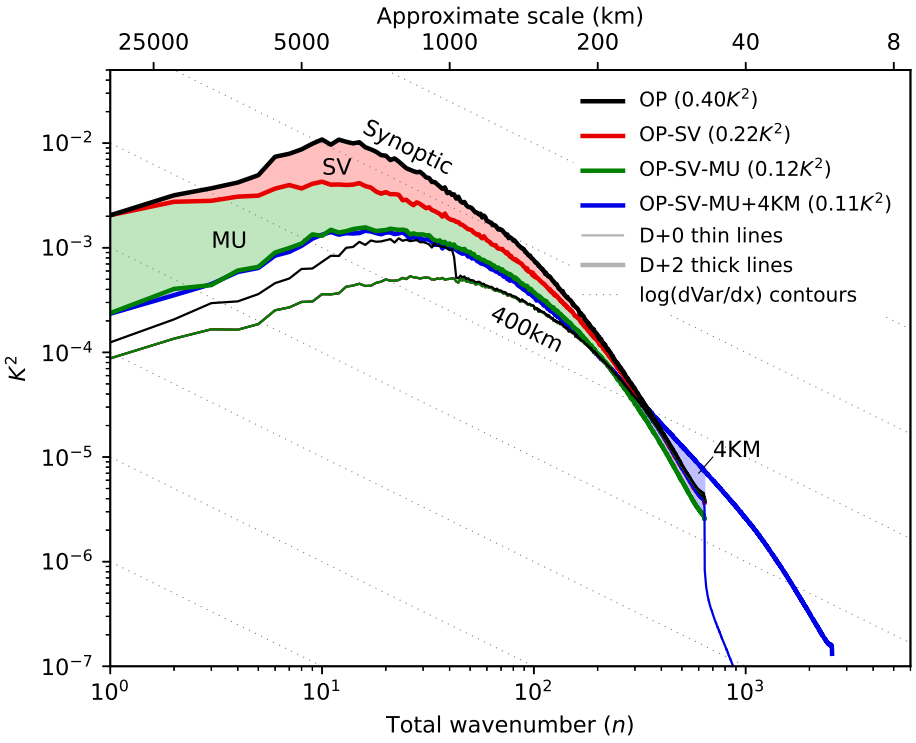

Figure B1. Spectra of ensemble variance for the indicated experiments, summarised in Fig. 9. Spectra show the variance of the ensemble filtered on total wavenumber n for global spherical harmonics. The approximate scale of each harmonic is indicated on the top x-axis (this relates to a half wavelength assuming equal zonal and meridional wavenumbers). Thin lines show initial time and thick lines show day 2. The red, green and blue shaded regions indicate the impacts at day 2 of including initial singular vector (SV) perturbations, model uncertainty (MU) representation and increasing the gridpoint resolution of the forecast model to ~ 4 km. Diagonal dotted lines are contours of variance contribution per linear unit length on the x-axis, with contour value indicated where they intersect the y-axis. Total variance for each experiment is indicated in parenthesis in the key.

scales. The thick blue curve shows the day 2 spectrum when the gridpoint resolution of the forecast model is increased to ~ 4 km. The spectrum is seen to 'shallow' even at scales already represented by the original ~ 16 km model (as indicated by the blue shaded region). A similar plot based on the same set of experiments for the second initial date (00 UTC on 17 January 2020) is almost indistinguishable from Fig. B1.

Author contributions. This study arose from close collaboration between the authors during HW's period as an ECMWF Fellow, and from ideas generated in the Warm Conveyor Belt workshop, which they co-organised in 2020. The mathematical content and much of the diagnostic work for this study was completed by MJR, with HW providing important guidance throughout.

Competing interests. The authors have no competing interests in this study

Acknowledgements. For the sensitivity experiments, the authors would like to thank Peter Bechtold, Andrew Dawson, Richard Forbes, Alan Geer, Elias Hólm, Bruce Ingleby, Simon Lang, Inna Polichtchouk, and Gabor Radnoti for their considerable. The authors would also like to thank Katharina Heitmann (ETH Zurich) for preparing Fig. 1 and Jonathan Day, Rebecca Emerton, David Lavers, Linus Magnusson, Florian Pappenberger, and David Richardson for useful discussions about this work. Finally, the authors would like to express their great appreciation to the two reviewers for their time and insightful remarks, which have led to improvements in this manuscript.

References

Ahmadi-Givi, F., Graig, G. C., and Plant, R. S.: The dynamics of a midlatitude cyclone with very strong latent-heat release, Quart. J. Roy. Meteor. Soc., 130, 295–323, https://doi.org/10.1256/qj.02.226, 2004.

Baumgart, M. and Riemer, M.: Processes governing the amplification of ensemble spread in a medium-range forecast with large forecast uncertainty, Quart. J. Roy. Meteor. Soc., 145, 3252–3270, https://doi.org/https://doi.org/10.1002/qj.3617, 2019.

Baumgart, M., Ghinassi, P., Wirth, V., Selz, T., Craig, G. C., and Riemer, M.: Quantitative View on the Processes Governing the Upscale Error Growth up to the Planetary Scale Using a Stochastic Convection Scheme, Mon. Wea. Rev., 147, 1713–1731, https://doi.org/10.1175/MWR-D-18-0292.1, 2019.

Bechtold, P., Köhler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, M. J., Vitart, F., and Balsamo, G.: Advances in simulating atmospheric variability with the ECMWF model: From synoptic to decadal time-scales, Quart. J. Roy. Meteor. Soc., 134, 1337–1351, 2008.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J., Beljaars, A., and Bormann, N.: Representing Equilibrium and Nonequilibrium Convection in Large-Scale Models, J. Atmos. Sci., 71, 734 – 753, https://doi.org/10.1175/JAS-D-13-0163.1, 2014.

Bishop, C. H., Etherton, B. J., and Majumdar, S. J.: Adaptive Sampling with the Ensemble Transform Kalman Filter. Part I: Theoretical Aspects, Mon. Wea. Rev., 129, 420 – 436, https://doi.org/10.1175/1520-0493(2001)129<0420:ASWTET>2.0.CO;2, 2001.

Bonavita, M., Hólm, E., Isaksen, L., and Fisher, M.: The evolution of the ECMWF hybrid data assimilation system, Quart. J. Roy. Meteor. Soc., 142, 287–303, https://doi.org/10.1002/qj.2652, 2016.

Buizza, R., Miller, M., and Palmer, T. N.: Stochastic representation of model uncertainties in the ECWMF Ensemble Prediction System, Quart. J. Roy. Meteor. Soc., 125, 2887–2908, https://doi.org/10.1002/qj.49712556006, 1999.

Chagnon, J. M., Gray, S. L., and Methven, J.: Diabatic processes modifying potential vorticity in a North Atlantic cyclone, Quart. J. Roy. Meteor. Soc., 139, 1270–1282, https://doi.org/https://doi.org/10.1002/qj.2037, 2013.

Christensen, H. M., Lock, S.-J., Moroz, I. M., and Palmer, T. N.: Introducing independent patterns into the Stochastically Perturbed Parametrization Tendencies (SPPT) scheme, Quart. J. Roy. Meteor. Soc., pp. n/a–n/a, https://doi.org/10.1002/qj.3075, 2017.

Dee, D. P.: Variational bias correction of radiance data in the ECMWF system., in: ECMWF Workshop on Assimilation of High Spectral Resolution Sounders in NWP, pp. 97–112, ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK, 2004.

Durran, D. R. and Gingrich, M.: Atmospheric Predictability: Why Butterflies Are Not of Practical Importance, J. Atmos. Sci., 71, 2476–2488, https://doi.org/10.1175/JAS-D-14-0007.1, 2014.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 10 143–10 162, https://doi.org/10.1029/94JC00572, 1994.

Geer, A. J., Lonitz, K., Weston, P., Kazumori, M., Okamoto, K., Zhu, Y., Liu, E. H., Collard, A., Bell, W., Migliorini, S., Chambon, P., Fourrié, N., Kim, M.-J., Köpken-Watts, C., and Schraff, C.: All-sky satellite data assimilation at operational weather forecasting centres, Quart. J. Roy. Meteor. Soc., 144, 1191–1217, https://doi.org/https://doi.org/10.1002/qj.3202, 2018.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, J. Am. Statist. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Grams, C. M. and Blumer, S. R.: European high-impact weather caused by the downstream response to the extratropical transition of North Atlantic Hurricane Katia (2011), Geophys. Res. Lett., 42, 8738–8748, https://doi.org/https://doi.org/10.1002/2015GL066253, 2015.

Grams, C. M., Magnusson, L., and Madonna, E.: An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study, Quart. J. Roy. Meteor. Soc., 144, 2577–2591, https://doi.org/https://doi.org/10.1002/qj.3353, 2018.

Hakim, G. J.: Role of Nonmodal Growth and Nonlinearity in Cyclogenesis Initial-Value Problems, J. Atmos. Sci., 57, 2951–2967, https://doi.org/10.1175/1520-0469(2000)057<2951:RONGAN>2.0.CO;2, 2000.

Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Mon. Wea. Rev., 129, 550–560, https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2, 2001.

Hartigan, J. A. and Wong, M. A.: Algorithm AS136: A K-means clustering algorithm, J. R. Stat. Soc., C: Appl. Stat., 28, 100–108, 1979.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quart. J. Roy. Meteor. Soc., 146, 1999–2049, https://doi.org/https://doi.org/10.1002/qj.3803, 2020.

Hirons, L. C., Inness, P., Vitart, F., and Bechtold, P.: Understanding advances in the simulation of intraseasonal variability in the ECMWF model. Part I: The representation of the MJO, Quart. J. Roy. Meteor. Soc., 139, 1417–1426, https://doi.org/https://doi.org/10.1002/qj.2060, 2013.

Holton, J. R.: An Introduction to Dynamic Meteorology, Academic Press, fouth edn., 553 pp., 2004.

Hoskins, B. J., McIntyre, M. E., and Robertson, A. W.: On the use and significance of isentropic potential vorticity maps, Quart. J. Roy. Meteor. Soc., 111, 877–946, https://doi.org/10.1002/qj.49711147002, 1985.

Hunt, B. R., Kostelich, E. J., and Szunyogh, I.: Efficient data assimilation for spatiotemporal chaos: A local ensemble transform Kalman filter, Physica D: Nonlinear Phenomena, 230, 112–126, https://doi.org/10.1016/j.physd.2006.11.008, 2007.

Isaksen, L., Hasler, J., Buizza, R., and Leutbecher, M.: The new Ensemble of Data Assimilations., ECMWF Newsletter 123, ECMWF, Shinfield Park, Reading, Berkshire, RG2 9AX, UK, available at http://www.ecmwf.int/publications, 2010.

Joos, H. and Forbes, R. M.: Impact of different IFS microphysics on a warm conveyor belt and the downstream flow evolution, Quart. J. Roy. Meteor. Soc., 142, 2727–2739, https://doi.org/https://doi.org/10.1002/qj.2863, 2016.

Judt, F.: Insights into Atmospheric Predictability through Global Convection-Permitting Model Simulations, J. Atmos. Sci., 75, 1477 – 1497, https://doi.org/10.1175/JAS-D-17-0343.1, 2018.

Laloyaux, P., Bonavita, M., Chrust, M., and Gürol, S.: Exploring the potential and limitations of weak-constraint 4D-Var, Quart. J. Roy. Meteor. Soc., 146, 4067–4082, https://doi.org/10.1002/qj.3891, 2020.

Lang, S. T. K., Dawson, A., Diamantakis, M., Dueben, P., Hatfield, S., Leutbecher, M., Palmer, T., Prates, F., Roberts, C. D., Sandu, I., and Wedi, N.: More accuracy with less precision, Quart. J. Roy. Meteor. Soc., n/a, https://doi.org/https://doi.org/10.1002/qj.4181, 2021.

Leutbecher, M. and Lang, S. T. K.: On the reliability of ensemble variance in subspaces defined by singular vectors, Quart. J. Roy. Meteor. Soc., 140, 1453–1466, https://doi.org/10.1002/qj.2229, 2014.

Leutbecher, M. and Palmer, T. N.: Ensemble Forecasting, J. Comp. Phys., 227, 3515–3539, https://doi.org/10.1016/j.jcp.2007.02.014, also available as ECMWF Tech. Memo. 514, 2008.

Lillo, S. P. and Parsons, D. B.: Investigating the dynamics of error growth in ECMWF medium-range forecast busts, Quart. J. Roy. Meteor. Soc., 143, 1211–1226, https://doi.org/10.1002/qj.2938, 2017.

Lorenz, E. N.: Deterministic non periodic flow, J. Atmos. Sci., 20, 130–141, https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2, 1963.

Lorenz, E. N.: Predictability; Does the flap of a Butterfly's wings in Brazil Set Off a Tornado in Texas?, Presented at the American Association for the Advancement of Science, 139th meeting, https://web.archive.org/web/20130612164541/http://eaps4.mit.edu/research/Lorenz/Butterfly_1972.pdf, archived 2013-06-12 at the Wayback Machine, Accessed: 2021-11-11, 1972.

Madonna, E., Wernli, H., Joos, H., and Martius, O.: Warm Conveyor Belts in the ERA-Interim Dataset (1979–2010). Part I: Climatology and Potential Vorticity Evolution, J. Climate, 27, 3–26, https://doi.org/10.1175/JCLI-D-12-00720.1, 2014.

Magnusson, L., Nycander, J., and Källén, E.: Flow-dependent versus flow-independent initial perturbations for ensemble prediction, Tellus A: Dynamic Meteorology and Oceanography, 61, 194–209, https://doi.org/10.1111/j.1600-0870.2008.00385.x, 2009.

McCabe, A., Swinbank, R., Tennant, W., and Lock, A.: Representing model uncertainty in the Met Office convection-permitting ensemble prediction system and its impact on fog forecasting, Quart. J. Roy. Meteor. Soc., 142, 2897–2910, https://doi.org/10.1002/qj.2876, 2016.

Molteni, F. and Palmer, T. N.: Predictability and finite-time instability of the northern winter circulation, Quart. J. Roy. Meteor. Soc., 119, 269–298, https://doi.org/10.1002/qj.49711951004, 1993.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and validation, Quart. J. Roy. Meteor. Soc., 122, 73–119, https://doi.org/10.1002/qj.49712252905, 1996.

Oertel, A., Boettcher, M., Joos, H., Sprenger, M., and Wernli, H.: Potential vorticity structure of embedded convection in a warm conveyor belt and its relevance for large-scale dynamics, Weather Clim. Dynam., 1, 127–153, https://doi.org/10.5194/wcd-1-127-2020, 2020.

Palmer, T., Döering, A., and Seregin, G.: The real butterfly effect, Nonlinearity, 27, R123–R141, https://doi.org/10.1088/0951-7715/27/9/r123, 2014.

Palmer, T. N., Molteni, F., Mureau, R., Buizza, R., Chapelet, P., and Tribbia, J.: Ensemble prediction, Tech. rep., ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK, available at http://www.ecmwf.int/publications, 1992.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F., and Simmons, A.: The ECMWF operational implementation of four-dimensional variational assimilation. I: Experimental results with simplified physics, Quart. J. Roy. Meteor. Soc., 126, 1143–1170, https://doi.org/10.1002/qj.49712656415, 2000.

Rennie, M. P., Isaksen, L., Weiler, F., de Kloe, J., Kanitz, T., and Reitebuch, O.: The impact of Aeolus wind retrievals on ECMWF global weather forecasts, Quart. J. Roy. Meteor. Soc., 147, 3555–3586, https://doi.org/https://doi.org/10.1002/qj.4142, 2021.

Riemer, M. and Jones, S. C.: Interaction of a tropical cyclone with a high-amplitude, midlatitude wave pattern: Waviness analysis, trough deformation and track bifurcation, Quart. J. Roy. Meteor. Soc., 140, 1362–1376, https://doi.org/https://doi.org/10.1002/qj.2221, 2014.

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., Diamantakis, M., Earnshaw, P., Garcia-Mendez, A., Isaksen, L., Källén, E., Klocke, D., Lopez, P., McNally, T., Persson, A., Prates, F., and Wedi, N.: Characteristics of occasional poor medium-range weather forecasts for Europe, Bull. Amer. Meteor. Soc., 94, 1393–1405, https://doi.org/10.1175/BAMS-D-12-00099.1, 2013.

Rodwell, M. J., Lang, S. T. K., Ingleby, N. B., Bormann, N., Hólm, E., Rabier, F., Richardson, D. S., and Yamaguchi, M.: Reliability in ensemble data assimilation, Quart. J. Roy. Meteor. Soc., 142, 443–454, https://doi.org/10.1002/qj.2663, 2016.

Rodwell, M. J., Richardson, D. S., Parsons, D. B., and Wernli, H.: Flow-Dependent Reliability: A Path to More Skillful Ensemble Forecasts, Bull. Amer. Meteor. Soc., 99, 1015–1026, https://doi.org/10.1175/BAMS-D-17-0027.1, 2018.

Rodwell, M. J., Hammond, J., Thornton, S., and Richardson, D. S.: User decisions, and how these could guide developments in probabilistic forecasting, Quart. J. Roy. Meteor. Soc., 146, 3266–3284, https://doi.org/10.1002/qj.3845, 2020.

Saetra, Ø., Hersbach, H., Bidlot, J. R., and Richardson, D. S.: Effects of Observation Errors on the Statistics for Ensemble Spread and Reliability, Mon. Wea. Rev., 132, 1487–1501, https://doi.org/10.1175/1520-0493(2004)132<1487:EOOEOT>2.0.CO;2, 2004.

Sanders, F.: The evaluation of subjective probability forecasts, Sci. Rept. 5, MIT, Dept. of Earth, Atmospheric and Planetary Sciences, 77 Massachusetts Avenue, Cambridge, MA 02139-4307, USA, 63 pp., 1958.

Sanders, F. and Hoskins, B. J.: An Easy Method for Estimation of Q-Vectors from Weather Maps, Wea. Forecasting, 5, 346 – 353, https://doi.org/10.1175/1520-0434(1990)005<0346:AEMFEO>2.0.CO;2, 1990.

Selz, T., Riemer, M., and Craig, G. C.: The Transition from Practical to Intrinsic Predictability of Midlatitude Weather, Journal of the Atmospheric Sciences, 79, 2013–2030, https://doi.org/10.1175/JAS-D-21-0271.1, 2022.

Shutts, G. J.: A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems, Tech. Rep. 449, ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, UK, 2004.

Sun, Y. Q. and Zhang, F.: Intrinsic versus Practical Limits of Atmospheric Predictability and the Significance of the Butterfly Effect, J. Atmos. Sci., 73, 1419–1438, https://doi.org/10.1175/JAS-D-15-0142.1, 2016.

Sutton, O. G.: The development of meteorology as an exact science, Quart. J. Roy. Meteor. Soc., 80, 328–338, https://doi.org/10.1002/qj.49708034503, 1954.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T. M., Hewson, T. D., Keller, J. H., Matsueda, M., Methven, J., Pappenberger, F., Scheuerer, M., Titley, H. A., Wilson, L., and Yamaguchi, M.: The TIGGE Project and Its Achievements, Bull. Amer. Meteor. Soc., 97, 49–67, https://doi.org/10.1175/BAMS-D-13-00191.1, 2016.

Tiedtke, M.: A Comprehensive Mass Flux Scheme for Cumulus Parameterization in Large-Scale Models, Mon. Wea. Rev., 117, 1779–1800, 1989.

Tribbia, J. J. and Baumhefner, D. P.: Scale Interactions and Atmospheric Predictability: An Updated Perspective, Mon. Wea. Rev., 132, 703–713, https://doi.org/10.1175/1520-0493(2004)132<0703:SIAAPA>2.0.CO;2, 2004.

Váňa, F., Düben, P., Lang, S., Palmer, T., Leutbecher, M., Salmond, D., and Carver, G.: Single Precision in Weather Forecasting Models: An Evaluation with the IFS, Mon. Wea. Rev., 145, 495–502, https://doi.org/10.1175/MWR-D-16-0228.1, 2017.

Wedi, N. P., Polichtchouk, I., Dueben, P., Anantharaj, V. G., Bauer, P., Boussetta, S., Browne, P., Deconinck, W., Gaudin, W., Hadade, I., Hatfield, S., Iffrig, O., Lopez, P., Maciel, P., Mueller, A., Saarinen, S., Sandu, I., Quintino, T., and Vitart, F.: A Baseline for Global Weather and Climate Simulations at 1 km Resolution, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 192, https://doi.org/10.1029/2020MS002192, 2020.

Wernli, H. and Davies, H. C.: A Lagrangian-based analysis of extratropical cyclones. I: The method and some applications, Quart. J. Roy. Meteor. Soc., 123, 467–489, https://doi.org/10.1002/qj.49712353811, 1997.

Wernli, H. and Schwierz, C.: Surface Cyclones in the ERA-40 Dataset (1958–2001). Part I: Novel Identification Method and Global Climatology, J. Atmos. Sci., 63, 2486 – 2507, https://doi.org/10.1175/JAS3766.1, 2006.