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.

Global numerical weather prediction is often limited by Lorenz–type "butterflies" in the flow. Here we think of these as local flow configurations associated with pronounced uncertainty growth–rates, as demonstrated in short–range ensemble predictions. Some of these configurations correspond to potential instabilities of the flow. Often, forecasts for severe downstream

5   weather only show an improvement in skill when these butterflies have passed. Here we focus on the "cyclogenesis butterfly" — associated with baroclinic and convective instabilities in the extratropics. One question addressed is how do different operational ensemble forecast systems (within the "TIGGE" archive) represent the associated uncertainty growth–rates? To test which might be "better", an extended spread–error equation is used to investigate how well these non–linear models maintain short–range statistical reliability during cyclogenesis. For the European Centre for Medium–Range Weather Forecasts

10   (ECMWF) ensemble, considerable short–range over–spread is found in the North Atlantic stormtrack during winter 2020/21 — representing a source of untapped predictability (this result appears to generalise to other stormtracks and at least one other model). Flow–type clustering demonstrates that this over–spread is directly associated with the representation of cyclogenesis. We attempt to quantify the contributions to the total spread in cyclogenesis cases from initial uncertainty (as derived from ensemble data assimilation), singular vector perturbations, and model uncertainty. At day 2, we find that up to 25% can

15   be associated with the singular vectors and up to 25% with model uncertainty. The over–spread suggests that reductions in forecast error over recent years would permit further development of these uncertainty aspects. The sensitivities of spread to resolution, the explicit representation of convection, and the assimilation of local observations provide additional insight for future development.

1 Introduction

20   In numerical weather prediction (NWP), ensembles comprising multiple forecasts are generally made (Lewis, 2005) to represent the chaotic growth of initial and model uncertainty (Sutton, 1954; Lorenz, 1963). The aim of these ensembles is to estimate the solution to the Liouville equation (Ehrendorfer, 1994), which describes the state–vector distribution of an infinite set of trajectories, as they diverge from their initial conditions on the system's attractor. The initial distribution for ensemble forecasts is generally obtained through data assimilation (Rabier et al., 2000; Humphreys et al., 2012), in particular ensemble data as-

25 similation (Isaksen et al., 2010) which aims to represent flow–dependent error covariances in the background or "first–guess" (Bonavita et al., 2016) and in the observations (Geer et al., 2018), as well as model uncertainty (MU; Buizza et al., 1999) and a model grid's lack of representativity of point observations (Janjić et al., 2018). In some cases (including at ECMWF) "singular vector" (SV; Molteni and Palmer, 1993) perturbations, which represent the most strongly growing modes over a finite time interval (in the linear regime), are applied to the distribution of initial conditions as a pragmatic means of boosting ensemble

30 divergence.

As implied by the existence of SVs, the divergence of trajectories within state–space is not uniform over the attractor, and weather predictability is often limited by the Lorenz–type "butterflies" in the flow (Lorenz, 1972). Here we think of these as local flow configurations where the chaotic and exponential growth–rate of uncertainty is particularly strong. Some of these configurations correspond to potential instabilities of the flow — baroclinic (Lillo and Parsons, 2017) and convective (Sun

35 and Zhang, 2016; Rodwell et al., 2018) for example. The extratropical transition of tropical cyclones (TCs) can lead to a strong growth of uncertainty (Riemer and Jones, 2014) with implications for severe weather downstream (Grams and Blumer, 2015). This is perhaps an example of a more general issue associated with the interaction of a cut–off vortex with an upper–tropospheric trough (Grams et al., 2018; Baumgart and Riemer, 2019). Sometimes the seeding of uncertainty growth can be traced further back in time by regressing the ensemble against itself at an earlier lead–time. For example, strong growth of

40 uncertainty in the ECMWF ensemble during the interaction of TC Humberto with an extratropical upper–tropospheric trough in 2019 appeared to be seeded with local uncertainties in the TC three days earlier (not shown). Strong uncertainty growth for Medicane Zorbas in 2018, associated with large uncertainties in the position of a potential vorticity streamer over the Mediterranean, has also been linked to the downstream propagation of uncertainties in the North Atlantic waveguide several days before (Portmann et al., 2020).

45 Although the location and structure of the early uncertainty growth will reflect the SVs of the flow (even when SV perturbations have not been explicitly applied), the magnitude of the growth–rate will depend on how strongly the uncertainty projects onto these SVs (Leutbecher and Lang, 2014). Beyond the linear regime, growth–rates might be different again. Operational ensembles may differ from the true growth–rate due to necessary approximations, although the stochastic representation of model uncertainty (Buizza et al., 1999) aims to account for these approximations (Christensen, 2020). Hence there is a need to

50 evaluate the state–dependent growth–rate of uncertainty within operational forecasts. Here the focus is on what is termed here the "cyclogenesis butterfly" — associated with the intensification of extratropical cyclones. The key question is "How well do models maintain statistical reliability during composites of cyclogenesis events?" The metric of reliability, discussed more fully in section 6, being an extension of the "spread–error" relationship: which states that the mean ensemble variance (Spread^2) should agree with the mean–squared error of the ensemble–mean (Error^2). In line with the examples discussed above, much

55 of the focus of this study is on reliability in the upper–troposphere, where uncertainties can propagate along the Jetstream / wave–guide and give rise to uncertainties in down–stream features such as anticyclonic "blocking" and extreme precipitation.

The study is structured as follows. Section 2 gives details of the data used. This includes data obtained for several of the world's operational ensemble forecasting systems from the "TIGGE" archive. Section 3 introduces two cyclogenesis case studies, and Section 4 discusses the quantification of uncertainty growth–rates. Section 5 then calculates these growth–rates

for the two case studies, and highlights differences between operational forecast systems in the TIGGE archive. To assess how well statistical reliability is maintained during such events, an "extended spread–error equation" is derived in Section 6. This accounts for bias and analysis uncertainty to permit the assessment of reliability at all lead–times. In Section 7 the extended spread–error equation is applied to two TIGGE models for the December 2020 – February 2021 (DJF 2020/21) season at a lead–time of two days. To differing extents, both models are shown to be "over–spread" in the region of the North Atlantic stormtrack. Focusing on the western part of the North Atlantic stormtrack, a clustering technique is used in Section 8 to identify cyclogenesis events in DJF 2020/21. Section 9 demonstrates that the seasonal–mean over–spread is associated mainly with these cyclogenesis events. Section 10 returns to the case studies from Section 3 to quantify the spread contributions from initial and model uncertainty, and to explore the sensitivity to resolution, the explicit representation of convection, and the assimilation of local observations. Conclusions and a discussion are presented in Section 11. The Appendix is used to provide some mathematical derivations, further detail, and additional figures. Supplementary material includes animations of uncertainty growth–rates.

2 Data

Here, a brief description of the datasets used is given. Further details are provided later in the text. The ECMWF re–analysis version 5 (ERA5; Hersbach et al., 2020) is used to show the analysed development of the two case studies. The background forecasts from the ECMWF Ensemble of Data Assimilations (EDA; Isaksen et al., 2010) are used to diagnose flow–dependent growth–rates in the standard deviation of (isentropic) potential vorticity (PV) for the ECMWF model. Further information on the EDA is presented in Appendix E. The International Grand Global Ensemble (TIGGE) archive (Swinbank et al., 2016) 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 source of data for diagnostic (and other) studies. The four centres focused on here, their archived ensemble sizes (perturbed + control) and data resolutions, are ECMWF (50+1mem, ∼15 km), the Japan Meteorological Agency (JMA, 26+1mem, ∼139 km), the United States' National Centers for Environmental Prediction (NCEP 30+1mem, ∼56 km), and the United Kingdom's Met Office (UKMO 17+1mem, ∼32 km). The 00 and 12 UTC initialisation times are used here for each model, although they are also available at other times for some centres. This data is used to compare flow–dependent growth–rates in the standard deviation of geopotential height amongst these four models, and for the systematic extended spread–error evaluation. Finally, a suite of sensitivity experiments is conducted with the ECMWF Integrated Forecast System (IFS) for the two case studies. This includes EDA and ensemble forecast (ENS) aspects. Further information on the IFS is presented in Appendix E.

3 Introducing two case studies of the cyclogenesis butterfly

Before a more systematic investigation of the "cyclogenesis butterfly", two cases are selected to highlight the salient features in its development and its uncertainty. The cases were chosen because they were quite "clean", without being strongly affected

by other flow perturbations in their environment, and because they involved both baroclinic and diabatic aspects. Figure 1 introduces the main synoptic–scale flow features of the two events. The cyclone of Case 1 develops on 26 November 2019 over the U.S. (not shown). By 18 UTC on 27 November it reaches the Great Lakes with a minimum sea level pressure (SLP) of about 990 hPa. The cyclone then reaches the North American east coast with a similar intensity (Fig. 1a). There is a cutoff in

95 PV on the 315 K isentrope ($PV\,315$) 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. The cyclone of Case 2 first occurs as a weak system with a SLP minimum larger than 1010 hPa over Ohio at 00 UTC on 16 January 2020 (not shown). It then propagates to the North Atlantic and deepens to less than 990 hPa ahead of a broad trough over eastern North America at 00

100 TC on 17 January, again with strong precipitation along the WCB ascent (Fig. 1c). During the next day, the cyclone deepens explosively and reaches a core pressure of about 960 hPa as it travels over the Atlantic south of Newfoundland (Fig. 1d). At that time the WCB extends from 25°N to 50°N, associated with intense precipitation, and the $PV\,315 = 2\,\mathrm{PV}$ unit ($2\,\mathrm{PVU}$) contour shows a prominent trough–ridge pattern over the western and central North Atlantic. This brief overview emphasizes the strong deepening of the two systems off the North American east coast and their baroclinic and diabatic character. Forecast

105 experiments that investigate the strong cyclogenesis / deepening periods for these two cases are initialised at 12 UTC on 28 November 2019 and 00 UTC on 17 January 2020.

4 The uncertainty growth–rate

The local uncertainty growth–rate of an ensemble forecast can be quantified as $\widehat{\sigma}^{-1}\,\partial\widehat{\sigma}/\partial t$ where $\widehat{\sigma}$ is the ensemble standard deviation of some 1–dimensional state–measure (of the atmosphere). Since much of this local growth–rate can be associated

110 with the advection of uncertainty, Rodwell et al. (2018) calculated a "Lagrangian" growth–rate $\widehat{\sigma}^{-1}(\partial\widehat{\sigma}/\partial t + \overline{\boldsymbol{v}} \cdot \boldsymbol{\nabla}\widehat{\sigma})$, which follows the ensemble–mean horizontal flow $\overline{\boldsymbol{v}}$. The state measure used by Rodwell et al. (2018) was isentropic potential vorticity P on some upper tropospheric isentrope θ. Following the derivation of Baumgart and Riemer (2019), but with a different formulation, the growth–rate of this measure can be related to sources of uncertainty growth via the equation

$$\frac{1}{\widehat{\sigma}}\left\{\frac{\partial\widehat{\sigma}}{\partial t} + \overline{\boldsymbol{v}} \cdot \boldsymbol{\nabla}_\theta\widehat{\sigma}\right\} = \frac{1}{\widehat{\sigma}^2}\overline{P'\left(NC - \boldsymbol{v}' \cdot \boldsymbol{\nabla}_\theta P\right)}\quad.$$
(1)

$$\underset{\text{Non–con}}{}\quad\underset{\text{Advective}}{}$$

115 Here, an overline denotes a mean over ensemble members, and a prime denotes an ensemble member's deviation from the mean. The "standard" (biased) estimator for variance is used throughout this study (with the bias explicitly removed where necessary) so that here $\widehat{\sigma}^2 = \overline{P'^2}$. The NC are the "non–conservative" tendencies in potential vorticity associated with diabatic and frictional processes for each ensemble member (see Appendix A for the derivation). Following the discussion of Baumgart and Riemer (2019), when Eq. (1) is applied to the upper troposphere, it shows how the growth–rate for PV will reflect local non–

120 conservative processes including near–tropopause radiative forcing (Chagnon et al., 2013), and non–linear advective tendencies
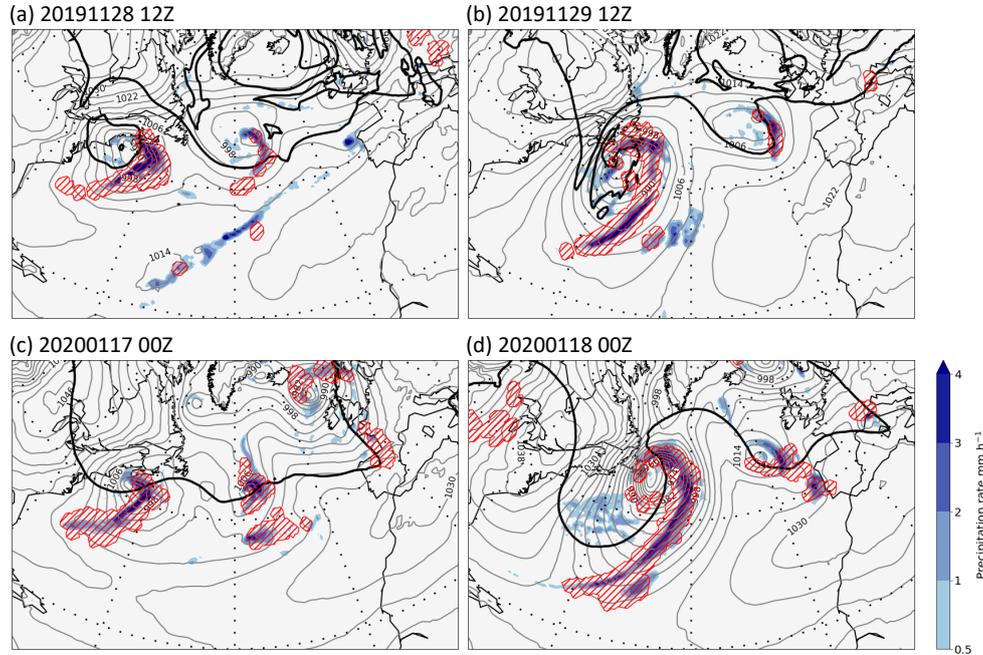
Figure 1. Synoptic overview, based on ERA5 reanalyses, of the two cyclogenesis events at (a,b) 12 UTC on 28 and 29 November 2019 for Case 1, and (c,d) 00 UTC on 17 and 18 January 2020 for Case 2. Shown are SLP (grey contours at intervals of 4 hPa), surface precipitation accumulated over the previous hour (colour shading, in mm h^{-1}), the $PV315 = 2$ 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 (e.g., Wernli and Davies, 1997; Madonna et al., 2014) and the ascent region (shown in red) corresponds to the envelope of the horizontal positions of all WCB trajectories that are located between 800 and 500 hPa at the indicated time.

due to e.g. upper tropospheric Rossby–wave dynamics (emphasized by Baumgart and Riemer, 2019), deep tropospheric interaction associated with baroclinic instability (e.g. Hoskins et al., 1985), and upper tropospheric divergence associated with dry balanced dynamics (as represented by the "Omega equation" and "Q–vectors"; Sanders and Hoskins, 1990) and mid–tropospheric diabatic forcing (e.g. Rodwell et al., 2013).

125   Note that there is a natural link between the growth–rate and the rate of change of the density of state–trajectories in the Liouville Equation (as discussed in section 1). Focusing on the probability density function (pdf) of any 1–dimensional state–measure, for small (initial) uncertainty and assuming the shape of the pdf does not change greatly over a short (∼12 h) forecast (only its mean and variance change), then $\widehat{\rho} \propto \widehat{\sigma}^{-1}$ where $\widehat{\rho}$ is an estimate of the relative liklihood of the ensemble mean (the value of the pdf at the ensemble–mean). Hence, for example, its rate of change following the ensemble–mean horizontal flow

130   can be estimated as minus the growth–rate, $\widehat{\rho}^{-1}(\partial\widehat{\rho}/\partial t + \boldsymbol{v} \cdot \boldsymbol{\nabla}\widehat{\rho}) = -\widehat{\sigma}^{-1}(\partial\widehat{\sigma}/\partial t + \boldsymbol{v} \cdot \boldsymbol{\nabla}\widehat{\sigma})$.
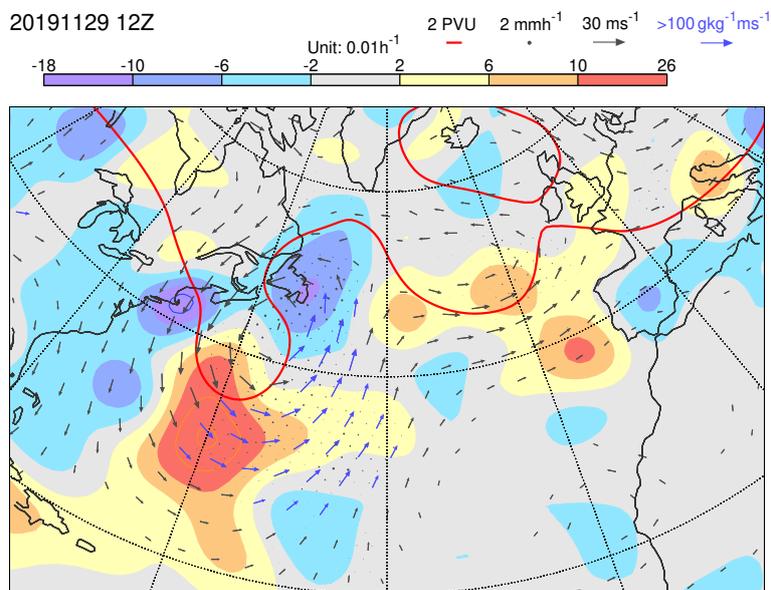
Figure 2. Single frame from the animation of cyclogenesis Case 1 in the EDA background ensemble forecast, focusing on synoptic spatio–temporal scales. The frame is centred at 12 UTC on 29 November 2019. Plotted are the control forecast $PV315 = 2$ PVU (red contour) and $\mathbf{v}850$ (vectors, which are plotted blue if the moisture flux at 850 hPa exceeds 100 $\text{gkg}^{-1}\text{ms}^{-1}$), the ensemble–mean precipitation (black dots, with radius proportional to the precipitation rate up to the maximum size at 2 mmh^{-1}), and the ensemble uncertainty growth–rate for the ensemble standard deviation of $PV315$ (shaded), as on the left–hand–side of Eq. (1). Contours extend the shading scheme to the most extreme values, which are indicated at the ends of the colour bar. Further details are available in the main text and in Appendix B.

5 Initial uncertainty growth–rates in the two case studies

Figure 2 and Fig. 3 show a single frame from animations of the two cyclogenesis events, one day into the strong cyclogenesis period. Shading depicts the $PV315$ uncertainty growth–rate following the ensemble–mean flow, the red contour shows where $PV315 = 2$ PVU and vectors display the 850 hPa horizontal wind field (coloured blue when the 850 hPa horizontal moisture flux exceeds 100 $\text{gkg}^{-1}\text{ms}^{-1}$), both from the control (unperturbed) forecast. Black dots indicate the ensemble–mean precip­itation rate. The animations are based on the ensemble of 12 h background forecasts made within the EDA for a sequence of analysis cycles. By concatenating these cycles and applying a "synoptic" spatio–temporal smoothing, the animations follow closely the true evolution of the synoptic flow, and highlight the features associated with enhanced growth–rates. Further details are in the caption to Fig. 2 and in the Appendix Section B. The animations of the case studies, and for the whole DJF 2020/21 season, are available in the supplementary material.

Strong growth–rates are seen at the southern extent of the upper–level trough in Fig. 2 and Fig. 3. More generally, in the animation for DJF 2020/21, they tend to occur just ahead of the base of the upper–level tough — often preceeding cyclogenesis. They also occur within strongly precipitating WCBs. Also evident in Fig. 2 and Fig. 3 are weaker negative growth–rates (where the ensemble is tending to converge) particularly in the ridge–building regions at the head of WCBs. This feature is repeated in
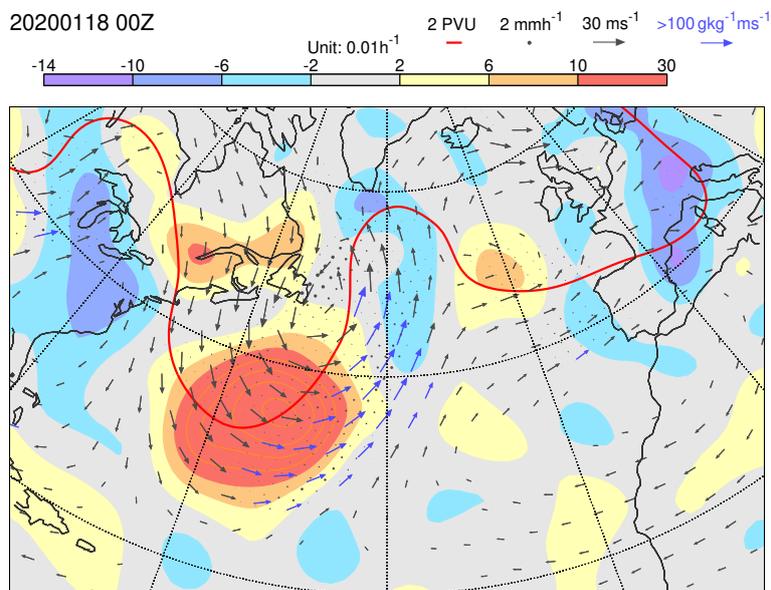
Figure 3. As Fig. 2 but for cyclogenesis Case 1, with the frame centred at 00 UTC on 18 January 2020.

numerous cases during the DJF 2020/21 season — a good example being around 2 February 2021. Further investigation of the potential processes discussed above, as in Baumgart and Riemer (2019), would be very valuable. For example, does differential vorticity advection (as in the Omega equation) play an important role in the region at the southern extent of the upper–level trough — either directly through uncertainty in the corresponding adiabatic ascent, or through uncertainties in the potential diabatic feedbacks? Moreover, how well do these modelled growth–rates correspond to the divergence of state trajectories on the real–world attractor? Here, a slightly easier question is first asked: "How well do growth–rates agree amongst the TIGGE models?"

Since potential vorticity is not archived in TIGGE, $Z250$ uncertainty growth–rates are calculated instead. Fig. 4 and Fig. 5 show the corresponding frames from animations based on the concatenation of 12 h forecasts from the four TIGGE models. For the ECMWF model, the $Z250$ growth–rates (panels a) there is reasonable agreement with the $PV315$ growth–rates in Fig. 2 and Fig. 3. The animation for the whole DJF 2021 season, also available in the supplementary material, displays numerous cyclogenesis events and suggests that all four models tend to enhance growth–rates at the southern extent of upper–level troughs. Nevertheless, differences between the models in their uncertainty growth–rates are very evident in Fig. 4 and Fig. 5, despite very similar upper–level height fields, low–level winds and precipitation. In particular, the ECMWF model tends to display the strongest growth–rates.

Clearly differences in initialisation procedures, differences in the representation of model uncertainty, and differences in the deterministic models themselves will all contribute to the growth–rate differences between the forecast systems. It would be useful to know which is closest to the truth, and why. In an attempt to begin to answer these questions, the next section
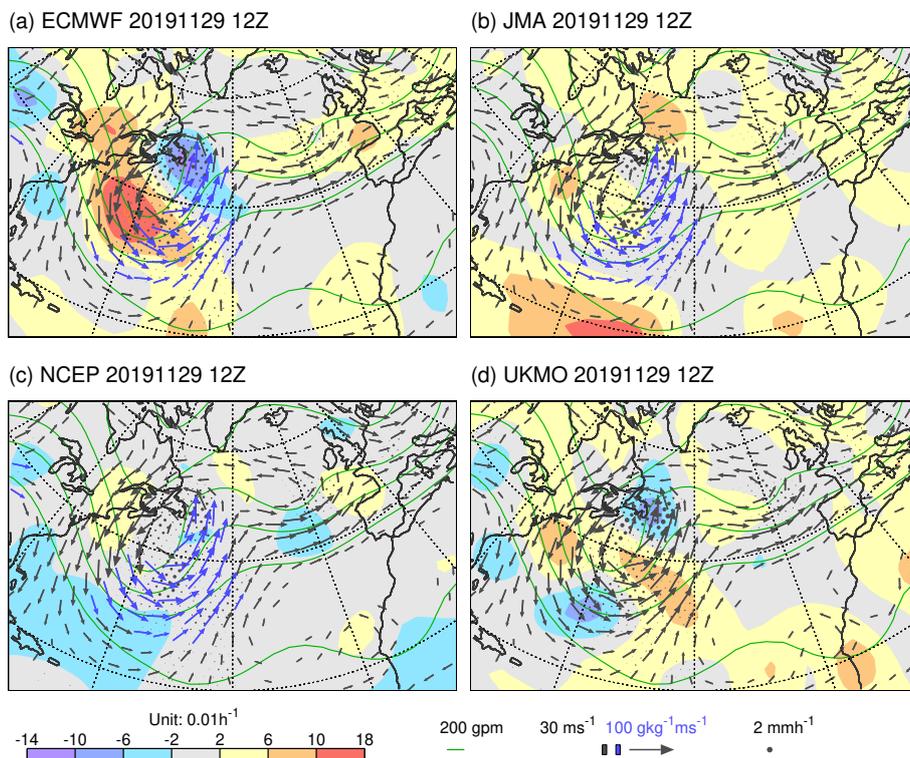
Figure 4. A single frame, centred at 12 UTC on 29 November 2019, from an animation of the cyclogenesis case 1, focusing on synoptic spatio–temporal scales, from (a) ECMWF, (b) JMA, (c) NCEP, and (d) UKMO ensemble forecast data obtained from the TIGGE archive. Plotted are the control forecast $Z250$ (green contours) and $\mathbf{v}850$ (vectors, which are plotted blue if the moisture flux at $850\,\text{hPa}$ exceeds $100\,\text{g}\text{kg}^{-1}\text{ms}^{-1}$), the ensemble–mean precipitation (black dots, with radius proportional to the precipitation rate up to the maximum size at $2\,\text{mmh}^{-1}$), and the ensemble uncertainty growth–rate for $Z250$ (shaded), as in Eq. (1). Contours extend the shading scheme to the most extreme values, which are indicated at the ends of the colour bar. Further details are available in the main text and in Appendix B.

discusses an extended spread–error equation which estimates deficiencies in short–range ensemble spread in the face of bias and analysis uncertainty.

165 6 The "extended spread–error" equation

For a reliable forecast system, the verifying truth should be statistically indistinguishable from any random sampling of the forecast distribution (Hamill, 2001; Saetra et al., 2004). A necessary condition of reliability is that, when averaged over many ensemble forecasts, the average difference between the truth and the ensemble mean (e.g., the mean–squared–error of the ensemble–mean) should match the average difference between an ensemble member and the ensemble mean (e.g., the ensemble

170 variance). This is the so–called "spread–error equation". In the present study, this equation is extended to account for bias
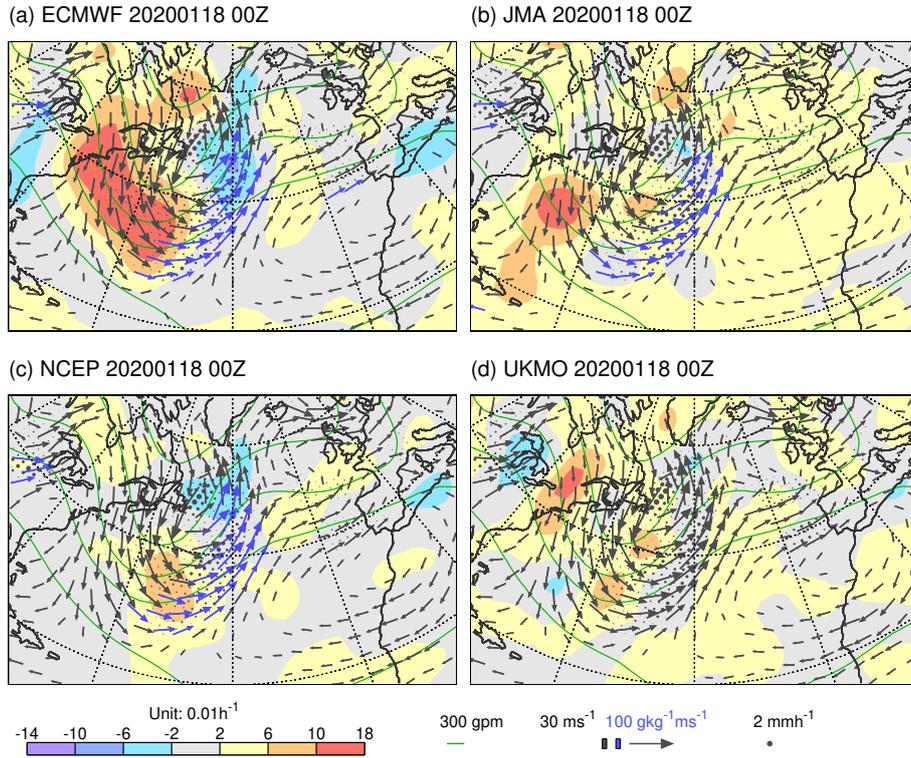
Figure 5. As Fig. 4, but centred at 0 UTC on 18 January 2020.

and analysis uncertainty (uncertainty in our knowledge of the truth), which can be non–negligible at the short lead–times considered, in an "extended spread–error equation":

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

(2)

An overline in Eq. (2) represents a mean over ensemble forecast start times. The schematic in Fig. 6 shows the salient aspects

175   of Eq. (2) for a single start time. An ensemble is considered to be a finite sampling of an "underlying distribution". The ensemble–mean forecast is shown as $\widehat{\mu}_{\widetilde{F}}$, where a hat indicates that this is an estimator of the mean $\mu_{\widetilde{F}}$ of the underlying forecast distribution, and a tilde indicates that the forecast is made with an imperfect model. Similarly, the ensemble–mean analysis $\widehat{\mu}_{\widetilde{A}}$ is an estimator of the mean $\mu_{\widetilde{A}}$ of the underlying analysis distribution. The line labelled d represents the departure of the ensemble–mean forecast from the ensemble–mean analysis $\widehat{\mu}_{\widetilde{F}} - \widehat{\mu}_{\widetilde{A}}$. The Error^2 term in Eq. (2) is the scaled mean–

180   square of these departures. (Scaling ensures that the equation is unbiased with a finite number of forecasts n and finite ensemble size m). The extended spread–error equation Eq. (2) aims to relate Error^2 to the sum of the mean–squares of the other lines in Fig. 6. This is derived fully in the Appendix C. The resulting Spread^2 and AnUnc^2 (analysis uncertainty) terms involve the
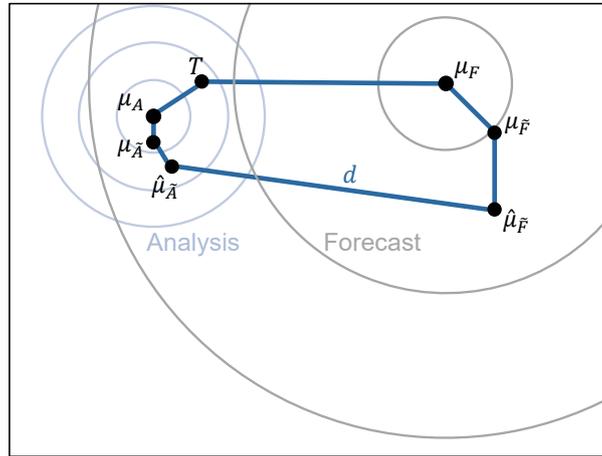
Figure 6. Schematic in 2 dimensions showing the "departure" d as the difference between the ensemble–mean of the operational forecast $\widehat{\mu}_{\tilde{F}}$ and the ensemble–mean of the verifying operational analysis $\widehat{\mu}_{\tilde{A}}$. The forecast and analysis ensembles can be considered as finite samplings of the underlying distributions with (mean,variance) = $(\mu_{\tilde{F}}, \sigma_{\tilde{F}}^2)$ and $(\mu_{\tilde{A}}, \sigma_{\tilde{A}}^2)$, respectively. Circles depict the hypothetical forecast and analysis distributions with (mean,variance) = (μ_F, σ_F^2) and (μ_A, σ_A^2), respectively, that would be created with a perfect model (and imperfect/incomplete observational information). The truth is indicated with a T.

variance estimators $\widehat{\sigma}_{\tilde{F}}^2$ and $\widehat{\sigma}_{\tilde{A}}^2$ of the forecast and analysis distributions, respectively. The Bias^2 is the squared–mean of the departures (the square of the estimated mean bias). The Residual$^{(2)}$ is the mean of the residuals that close the budget for each
185 initial time (assuming constant bias). The superscript $^{(2)}$ signifies that this is also in squared units, although it can be negative as well as positive.

If the mean Bias $\overline{d}$ is different from zero at some level of significance (here 5% with a t–test), then this indicates a bias in the mean of the underlying forecast distributions relative to those of a perfect model $(\mu_{\tilde{F}} - \mu_F)$ or a bias in the mean of the underlying analysis distributions relative to those made using a perfect model $(\mu_{\tilde{A}} - \mu_A)$. Similarly, if the Residual$^{(2)}$ is
190 different from zero at some level of significance, then this generally indicates a deficiency in the variance of the underlying forecast and/or analysis distribution. (A negative Residual$^{(2)}$ indicates a surplus in variance while positive values *can* indicate a deficit in variance provided the main additional term in the Residual$^{(2)}$ — 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
195 forecast distribution to begin to dominate.

For display purposes the terms in Eq. (2) can be put into more understandable units by taking their square–roots — signified by removing the superscript 2 label. Since $\overline{R}$ can be positive or negative, "Residual" = $\sqrt{|\overline{R}|} \cdot \text{SGN}(\overline{R})$ is plotted to retain the sign. One should be aware that the square–rooting will tend to make smaller terms look more important than they really are in the proper (squared) budget. Statistical significance is determined from the un–rooted terms (except for Bias^2, which must be
200 determined in its non–squared form).

As an aside, a very similar equation was presented by Rodwell et al. (2016) for application in "observation space". That equation represents a more fundamental "ideal" where observation error variances are assigned to represent their true values. In reality, these 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

205 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.

7 Seasonal–means of the extended spread–error equation

Figure 7 shows the (square–rooted) terms of the extended spread–error equation for $Z250$ based on all day–2 ensemble fore-

210 casts verifying in DJF 2020/21 from the ECMWF (top) and UK Met Office (middle) ensembles, and the differences (bottom). Focusing first on ECMWF (top), the North Atlantic winter stormtrack is evident as a region of enhanced ensemble spread (Fig. 7b). Even without the AnUnc and Bias contributions, the Spread is larger than required to balance the Error (Fig. 7a) — signifying "over–spread" in the stormtrack at day 2. Analysis uncertainty is also enhanced in the stormtrack region (Fig. 7c) and a statistically significant Bias is seen over the east coast of North America. Confirmation of the over–spread is seen with the

215 large negative Residual (Fig. 7e). 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. (2) is roughly $15^2 = 18^2 + 3^2 + 6^2 - 12^2 \, \mathrm{m}^2$. Because AnUnc and Bias are not negligible here, the reliable Spread (i.e., Spread in case of zero Residual) in this region would actually be $\sim 13.4^2 \, \mathrm{m}^2$, somewhat less than the $15^2 \, \mathrm{m}^2$ 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

220 Spread a little. In contrast, there is a positive Residual over the subtropical North Atlantic (Fig. 7e). Whether this indicates insufficient Spread to account for the elevated Errors in this region (Fig. 7a) or is associated with variance in forecast bias is also discussed later.

The corresponding budget for the UKMO (middle row of panels) also indicates an over–spread in the stormtrack at day 2 (Fig. 7j), but by a smaller degree. The difference plots (bottom row of panels) show that Error, AnUnc and |Bias| are all

225 generally smaller for ECMWF (Fig. 7k, Fig. 7m, Fig. 7n). In particular, ECMWF starts from a somewhat sharper analysis distribution outside the stormtrack region. However, ECMWF forecast Spread grows more rapidly in the stormtrack, and results in a stronger |Residual| than for the UKMO (Fig. 7o).

Note that the over–spread aspect is not specific to $Z250$ or the North Atlantic. Supplementary Figure D1 and Fig. D2 show that similar conclusions can be drawn for the more commonly verified field of geopotential heights at 500 hPa in all the

230 northern and southern hemispheric stormtracks in DJF 2020/21 and June – August 2021 (JJA 2021). The UKMO ensemble appears particularly over–spread in polar regions. Note that these supplementary figures show the non–rooted terms to give a better impression of the relative contributions to the overall budget in Eq. (2).
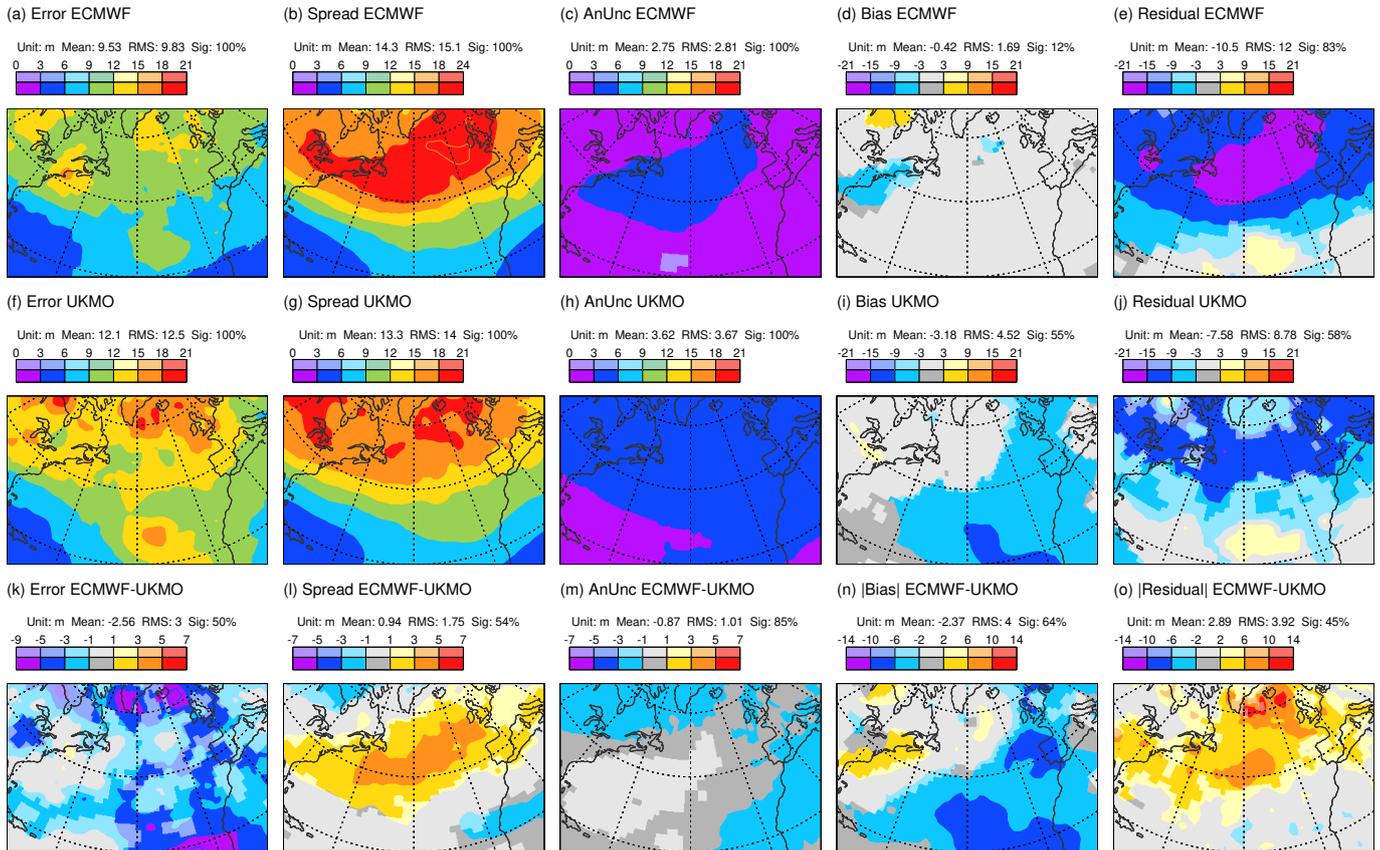
Figure 7. Square–roots of terms in the extended spread–error budget Eq. (2) for $Z250$ based on all day 2 forecasts verifying in DJF 2020/21. Note that Residual $= \sqrt{|R|}\mathrm{SGN}(R)$ to retain the correct sign. Data comes from the TIGGE archive for the ECMWF (top) and UKMO (middle) models, with the differences plotted on the bottom row. Note that the difference of absolute values is shown for Bias and Residual so that blue (red) implies that ECMWF is better (worse) than UKMO for any attribute difference. Contours extend the shading scheme to the most extreme values, which are indicated at the ends of the colour bars. 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). To avoid the potential for artificial weakening of significance associated with diurnal variability, the terms are diurnally averaged (average over 00 and 12 UTC) prior to testing.

A key question for this study is whether the ECMWF Residual (Fig. 7e) is associated with a general level of over–spread (e.g. associated with uncertainties in large–scale waves Durran and Gingrich, 2014), or whether it is associated with difficulties
235 in modelling particular flow evolutions, such as cyclogenesis? This question will be addressed based on a clustering of the flow configurations in DJF 2020/21. In the next section, the method of clustering is discussed.

8 Flow–type clustering

The method of K–means clustering (Hartigan and Wong, 1979) is used to obtain three flow clusters in a prescribed region. The data used within the clustering come from the ECMWF forecasts initialized at 00 and 12 UTC, with $Z250$, and zonal and

240 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. Three clusters should provide sufficient degrees of freedom to differentiate the possible flow–types. Each field is on a regular F32 (∼300 km) Gaussian grid, standardised (about its area– and temporal–mean)

245 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, and 100 such initialisations are performed. The minimum total within–cluster variance is generally found within the first 5 such initialisations — indicating that the optimal clusters have been identified.

250 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 warm conveyor belt activity (Madonna et al., 2014), and its size corresponds to a half–wavelength of a typical baroclinic wave. Figure 8 shows the same fields as in Fig. 4 and Fig. 5 but averaged over the three sets of date/times obtained from the K–means clustering in this (indicated) region. Cluster 1

255 (Fig. 8a, 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. Strong uncertainty growth–rate at the southern extent of the upper–level trough is also evident. (Note that the growth–rate is not used within the clustering algorithm since this is what will be evaluated). Cluster 2 (Fig. 8b, 75 date/times) shows a broader trough, weaker moisture flux, and possible cyclogenesis, displaced further downstream. Cluster 3 (Fig. 8c, 73 date/times) shows a diffuse ridge with a trough

260 even further downstream, and a surface anticyclone in the subtropical western North Atlantic.

A second clustering region is displaced north–eastward along the stormtrack [65°–35°W, 35°–55°N] and also contains 11×7 data points. Clustering results for this region are shown Fig. 9. Cluster 1 (Fig. 9a, 62 date/times) highlights further cyclogenesis with a closed cluster–mean circulation at 850 hPa over Newfoundland and with strong growth–rates.

In total, 91 date/times were identified as cyclogenesis flow–types (32+62 minus 3 duplicates). For each of these date/times

265 (and their 180-91=89 counterpart date/times), visual inspection of plots similar to those for the two case studies (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.
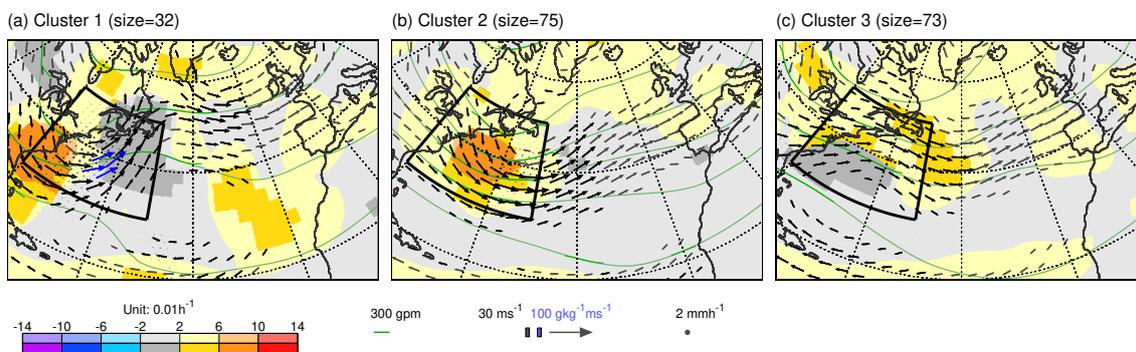
Figure 8. Means over the date/times for each cluster obtained from *K*–means clustering in the black–bordered region on the fields of $Z250$ (contours), $u850$ and $v850$ (vectors) from ECMWF control (unperturbed) forecasts at step 0, and ensemble–mean 12 h accumulated precipitation (dots) valid (or end of accumulation) at 00 and 12 UTC within the period DJF 2020/21. Although not used in the clustering, shading shows the corresponding mean 12 h uncertainty growth–rate Eq. (1) for $Z250$. Vectors are coloured blue when the humidity flux at 850 hPa exceeds $100\,\mathrm{g\,kg}^{-1}$. More saturated shading, contours and vectors indicate statistical significance at the 5% level (not accounting for autocorrelation due to the discontinuous nature of date/times in each cluster).
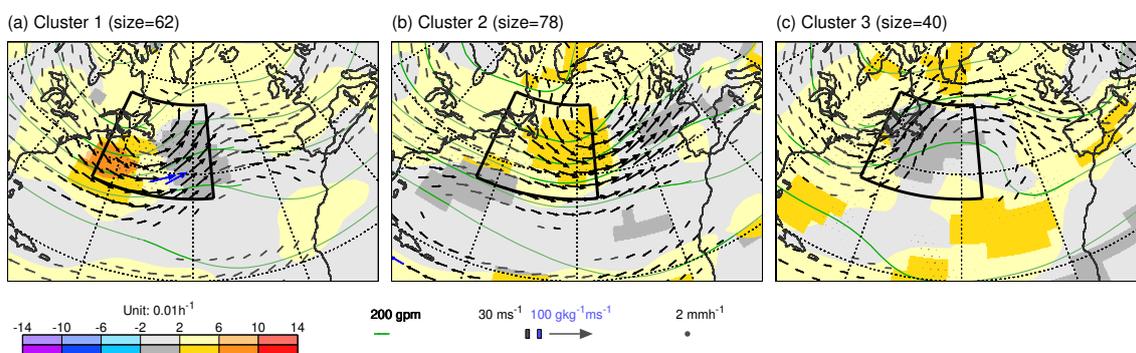
Figure 9. As Fig. 8 but for clustering in the second region (again indicated by a black border).

9 The extended spread–error equation for cyclogenesis flow–types

Because Fig. 8a and Fig. 9a 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 head 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 combined 91+89=180 date/times of the cyclogenesis and counterpart composites represent the same initial date/times used in Fig. 7 (top row).
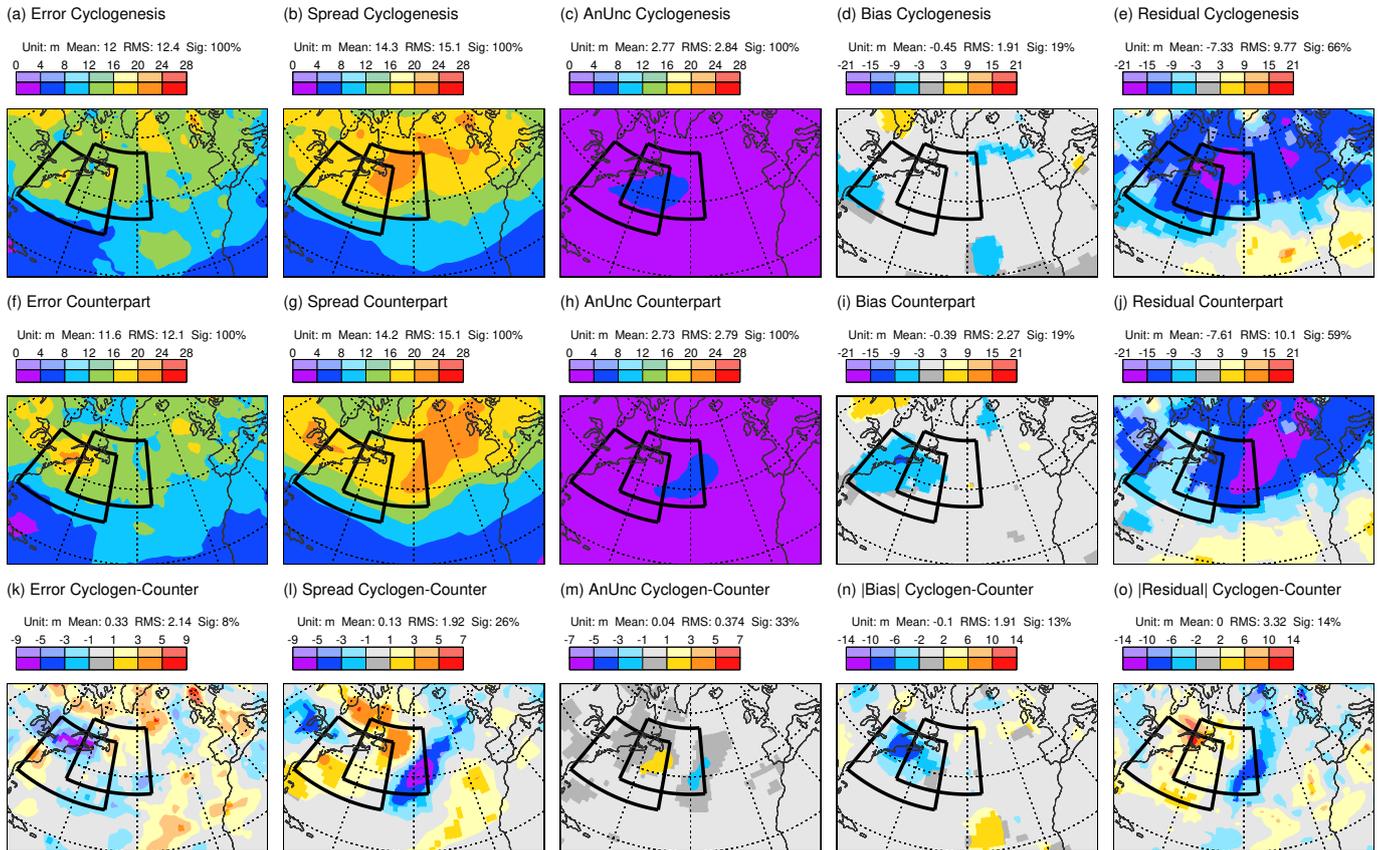
Figure 10. As Fig. 7 but separately for the cyclogenesis cluster date/times (top) and the non–cyclogenesis counterpart cluster date/times (middle), for forecasts that start two days prior to the cluster classifications. The bottom row shows cyclogenesis minus counterpart. The two clustering regions are indicated by black borders. Due to the discontinuity of date/times, diurnal averaging prior to statistical testing and the accounting for autocorrelation are not appropriate.

Figure 10 shows the (square–roots of) the terms in the extended spread–error equation Eq. (2) for $Z250$ at day 2, separately for cyclogenesis (top) and counterpart (middle) composites, and their difference (bottom). Comparison shows that ensemble spread for the cyclogenesis composite (Fig. 10b) is enhanced around the head of the North Atlantic stormtrack while the
280   spread for the counterpart composite (Fig. 10g) is centred more downstream. There are also corresponding differences in analysis uncertainty (Fig. 10c and Fig. 10h), possibly due to differing uncertainty growth–rates in the background forecasts used within the ensemble data assimilation process.

Figure 10n 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 $\sim 3^2 \, \text{m}^2$ — suggesting that this term is as important in (the
285   Residual term of) the day 2 extended spread–error budget for the whole of DJF 2020/21 as the explicitly represented analysis uncertainty (Fig. 7c), and would suggest an even stronger over–spread issue (cf. term estimates in Section 7). In contrast, the

positive Residual term over the subtropical North Atlantic (Fig. 7e) could reflect variance in forecast bias (∼Fig. 10n) rather than simply indicating ensemble under–spread. Thinking more broadly, it is possible that the variance in forecast bias might also be important for 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. 10d and Fig. 10i. Notice that the bias along the east coast of North America for the counterpart composite (Fig. 10i) appears to account for the increased errors seen in this region (compare Fig. 10f and Fig. 10a).

The overall assessment of ensemble spread is seen in the residual terms (Fig. 10e and Fig. 10j). Here it is evident that almost the entire over–spread in the region of focus, at the head of the North Atlantic winter stormtrack (Fig. 7e) is associated with the cyclogenesis composite. This is a key conclusion of this study.

The approach of this study was to assess reliability within a single season, because one instance of any given season is often all that is available within the operational forecast development process. One caveat to the approach taken is the lack of independent assessment; we are choosing initial conditions which we know do evolve with cyclogenesis. If anything, through "regression to the mean", this would suggest larger errors for the cyclogenesis composite at the head of the stormtrack — something which is not seen (cf. Fig. 10a and Fig. 10f). Hence, we do not feel this aspect compromises the conclusions drawn here.

10 Sensitivity experiments — including quantifying the roles of initial and model uncertainty

The final part of this paper returns to the two case studies of cyclogenesis discussed in Sections 3 and 5. Here, sensitivity studies are conducted to quantify the sources of uncertainty at day 2 in ECMWF's Integrated Forecasting System (IFS). The base configuration is IFS cycle 46r1, which was operational from 11 June 2019 to 20 June 2020. This includes the Ensemble of Data Assimilations (EDA), which combines observational and background information to provide the initial conditions for the ensemble forecast (ENS). Both the EDA and ENS are run at ∼18 km resolution with 50 members. For the interested reader, the salient details of IFS cycle 46r1, along with the minor technical changes required to run the sensitivity experiments, are presented in Appendix E.

The configuration diagram in Fig. 11 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-DC), 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 (as indicated by the vertical arrows in Fig. 11) allow the diagnosis of individual aspects. The conclusions are not thought to be sensitive to the ordering of the various modifications.

Figure 12 shows day 2 results for the standard deviation (spread) in $Z250$ from sensitivity experiments for cyclogenesis Case 1, 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. 4a). Grey contours show mean sea–level pressure, and the black contour shows where
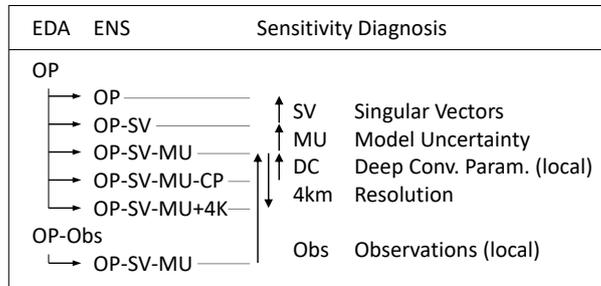
Figure 11. Configuration of IFS sensitivity experiments. These involve the Ensemble of Data Assimilations (EDA) and the Ensemble forecast (ENS). Differences between these configurations allow the diagnosis of sensitivity to individual aspects as indicated. See main text for further details.

$PV315 = 2\,\mathrm{PVU}$ — highlighting the location of the tropopause on the 315 K isentrope) in the unperturbed EDA analysis. A parallel set of results for spread in $PV315$ is shown in Fig. 13 (discussed later).

Figure 12a shows the OP configuration with a well–developed surface low pressure system, as discussed in relation to Fig. 1. The prominent warm conveyor belt (WCB) associated with this cyclone is seen to lead to the development of a prominent downstream upper–level ridge. The $Z250$ spread represents the temporal integral of the local tendency, and hence the effects of "material" generation and advection. The impact of the initial SV perturbations on $Z250$ spread (Fig. 12b) is particularly pronounced in the cyclone's cold sector and along the downstream tropopause. There are places where the SV impact on spread is half the total (so that the fraction of variance explained reaches 25%). The impact of the model uncertainty (MU) representation (Fig. 12c is particularly pronounced in the cyclone centre and in the region of the WCB ahead of the surface low. 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.

The impact of including parametrized deep convection (DC) in the indicated region (Fig. 12d; note the smaller contour interval) is a reduction in spread — particularly in the WCB region — that would otherwise be created when the model attempts to resolve this convection. 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 $\sim4\,\mathrm{km}$ is mixed. For $Z250$ spread (Fig. 12e; note the smaller contour interval), the impact is generally weak. In contrast the impact on $PV315$ spread (Fig. 13e) is strong, particularly within the WCB region. The smaller scales associated with $PV315$, and its sensitivity to vertical gradients in diabatic heating in the upper troposphere might help explain this sensitivity to the increase in resolution. 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. 13e and Fig. 13c). Forcing the model to explicitly resolve this convection (even if at the wrong $\sim18\,\mathrm{km}$ scale) appears to better locate the uncertainty with that generated by the $\sim4\,\mathrm{km}$ km model (cf. Fig. 13e and
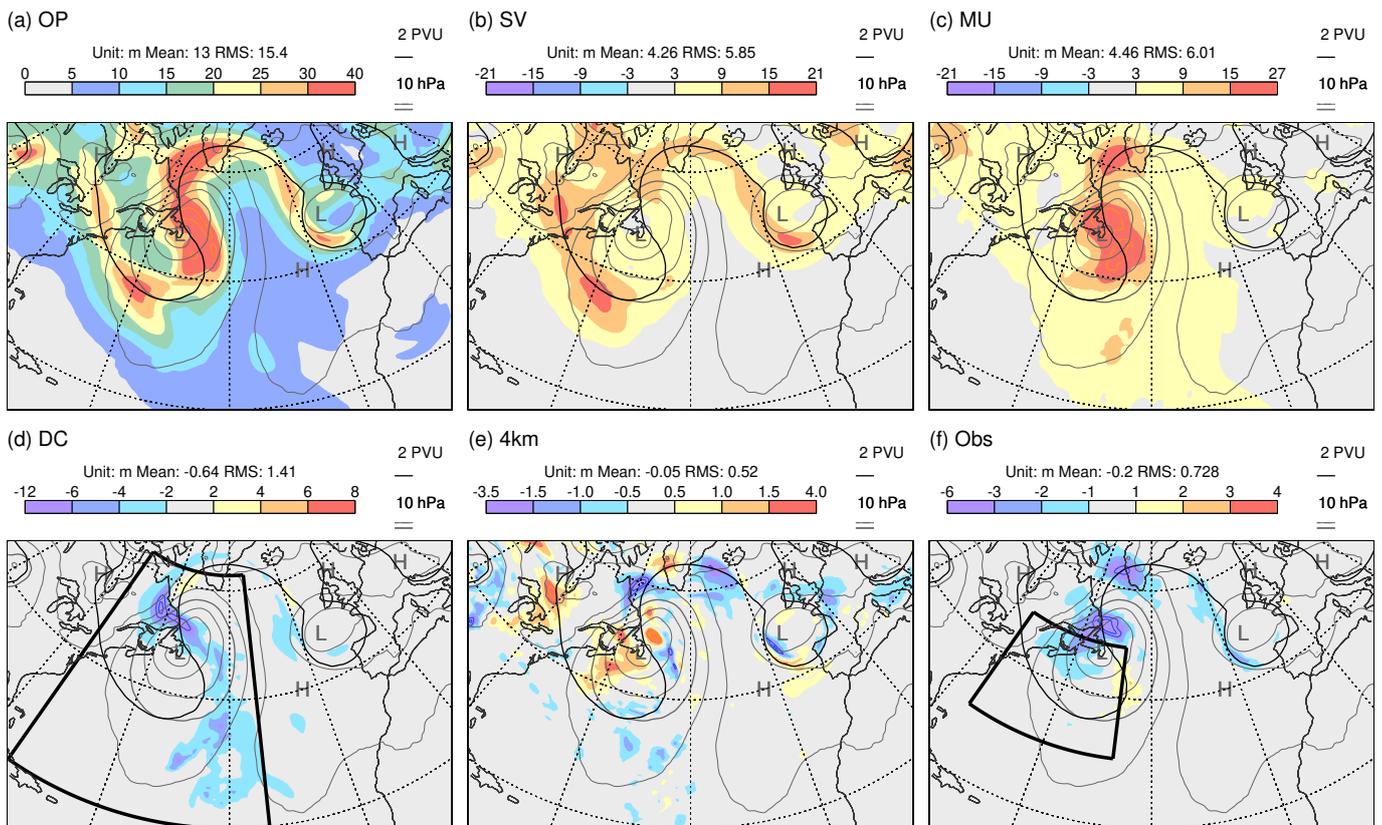
Figure 12. Sensitivity results for cyclogenesis Case 1, showing day–2 $Z250$ spread (shaded) from ECMWF ensemble forecast experiments initialised at 12 UTC on 28 November 2019. (a) Total spread for the near–operational configuration OP. (b)–(f) Differences in spread between experiments which highlight 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, (e) an increase in model grid resolution to ∼4 km, and (f) the assimilation of observations in the indicated region. Contours extend the shading scheme to the most extreme values, which are indicated at the ends of the colour bars. Also shown in each panel are the mean–sea–level pressure (grey contours) and PV=2 PVU on the 315 K isentrope (black contour) from the unperturbed EDA analysis.
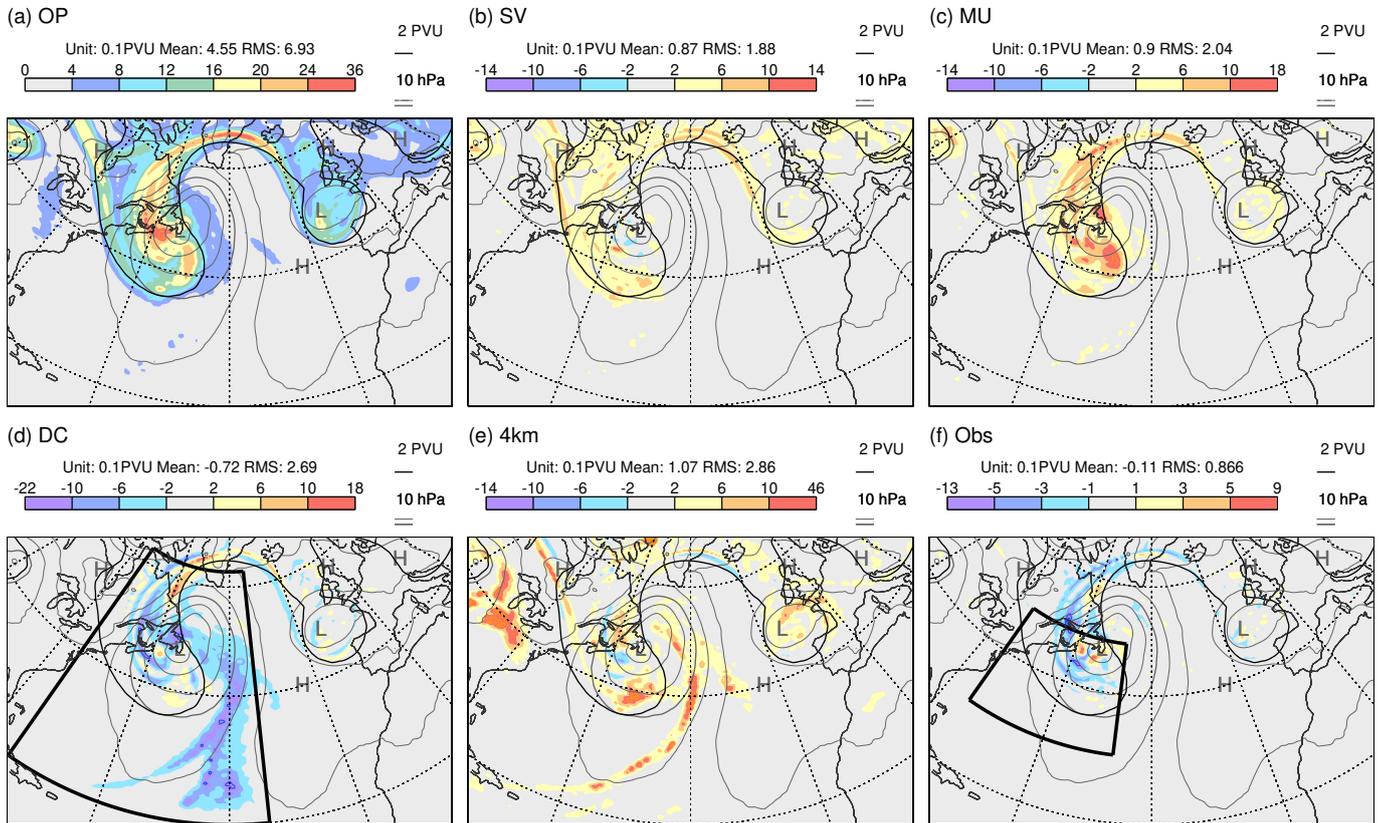
Figure 13. As Fig. 12, but shading shows the spread in $PV315$.

minus Fig. 13d). The increase in resolution also results in a small shift of ensemble–mean precipitation from parametrized to
345   resolved, with little change in the total (not shown).

Figure 12f shows the impact of using observations in the indicated region during the single data assimilation cycle that
generated the initial conditions (as shown in Fig. 11, this impact is deduced from a data denial experiment; observations
outside the indicated region are used in both experiments). The assimilation of observations in this baroclinic region acts to
reduce initial uncertainty (not shown) and to reduce day 2 spread in the developing low (Fig. 12f, Fig. 13f). This reduction in
350   spread is interesting because it helps quantify the role of local observations relative to that of remote observational information,
which can propagate into the region within the 4DVar process.

Figure 14 and Fig. 15 show the corresponding results for $Z250$ and $PV315$, respectively, for the cyclogenesis Case 2,
initialised at 00 UTC on 17 January 2020. Many results are similar — 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 (Fig. 14b) and MU (Fig. 14c),
355   with the SV impact largely in the cold sector and the MU impact in the region of the WCB. Again, the parametrization of
deep convection acts to reduce spread in the WCB (Fig. 14d). For $PV315$, the impact of increased resolution (Fig. 15e) is
again more similar to the effect of turning–off convective parametrization (minus Fig. 15d) than to that of model uncertainty
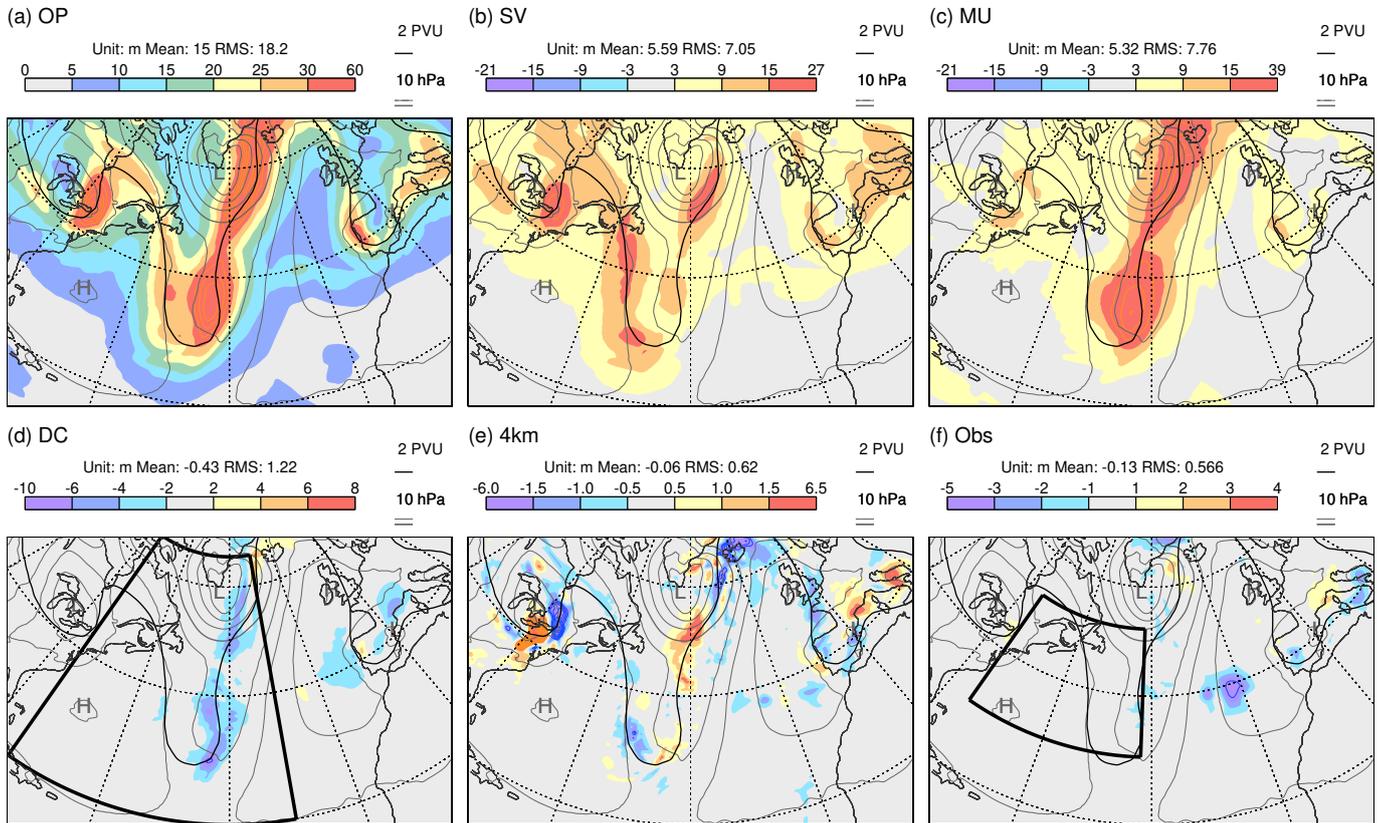
Figure 14. As Fig. 12 but for Cyclogenesis Case 2, with ENS forecast experiments initialised at 00 UTC on 17 January 2020.

(Fig. 15c). There is a reduction in initial spread from the assimilation of local observations (not shown), although it is weaker than in Case 1, and its impact at day 2 (Fig. 14f and Fig. 15f) is marginal.

11 Conclusions and Discussion

The "calibration" of ensemble spread to achieve reliability in an area–mean, temporal–mean sense can only take NWP development so far. To go further, it seems important to ensure that ensembles can maintain (short–range) reliability through their development of different flow–types. This study has focused on the flow–type associated with extratropical cyclogenesis. It has been demonstrated that the growth–rate of ensemble spread in 250 hPa geopotential heights within cyclogenesis varies substantially amongst four forecast models in the "TIGGE" archive. In theory, the growth–rate aims to represent (minus) the rate of change of probability density on the real–world's attractor — as described by the Liouville equation — so there should be a "correct" solution in terms of uncertainty growth–rate. While necessary approximations mean that all operational models will be compromised, it would be useful to have a framework for comparison, for deciding which model, or model update, is better able to maintain short–range reliability. Here, an approach based on short–range flow–dependent evaluation of an "extended
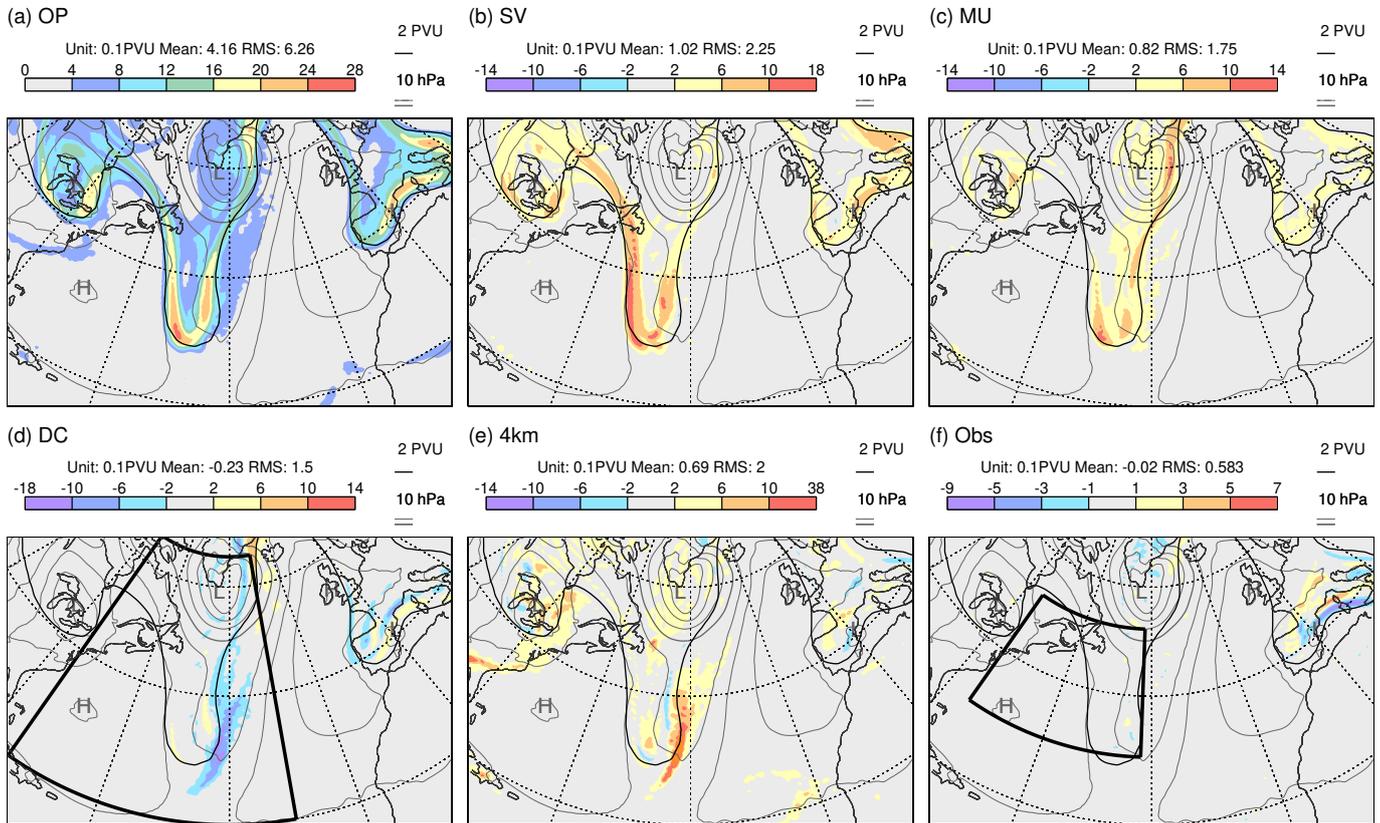
Figure 15. As Fig. 14, but shading shows the spread in $PV315$.

spread–error equation" has been developed. It is found that the two models evaluated in this way (from ECMWF and the UK's Met Office) are over–spread in the winter North Atlantic stormtrack (and indeed in all stormtracks). This represents a source of "untapped predictability" if it can be isolated and improved. Rather than representing a general over–spread in forecasts, it is demonstrated (for the ECMWF model over the western North Atlantic) to be associated with the cyclogenesis flow–type — involving an upper–level trough, and baroclinic and convective instabilities in the extra–tropics.

Case studies reveal the sources of uncertainty. A large part is associated with the chaotic growth of initial uncertainty within the deterministic model, but singular vector perturbations to the initial conditions and model uncertainty representations are also important. Dry singular vectors, which target baroclinic instabilities (Molteni and Palmer, 1993), and which are optimised at ECMWF for 2–day growth, may well be implicated in the over–spread. If so, 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 of model uncertainty. Furthermore, sensitivities to switching–off the parametrization of deep convection and to increasing model resolution suggest that the model uncertainty representation could be more strongly focused on convective instabilities (e.g., Christensen et al., 2017). One might speculate whether the effects of moist singular vectors could be represented within the formulation of model uncertainty itself.

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

code available

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

Appendix A: Derivation of local growth–rate equation for isentropic PV

If P_i, $\boldsymbol{v}_i$ are the isentropic potential vorticity and horizontal wind vector for ensemble member i ($=1,\ldots,m$), respectively,

390   then $\partial P_i/\partial t + \boldsymbol{v}_i \cdot \boldsymbol{\nabla}_\theta P_i = NC_i$, where t is time, $\boldsymbol{\nabla}_\theta$ is the gradient operator on the isentropic surface θ, and NC_i represents
the effects of "non–conservative" diabatic and frictional processes. Using an overline to denote a mean $\overline{X} = m^{-1} \sum_i X_i$, and
a prime to denote deviations from the mean $X_i' = X_i - \overline{X}$, then the variance of the P_i can be written as $\widehat{\sigma} = \overline{P'^2}$, and the local
growth–rate of the standard deviation written as

$$
\begin{aligned}
\frac{1}{\widehat{\sigma}} \frac{\partial \widehat{\sigma}}{\partial t} &= \frac{1}{2\widehat{\sigma}^2} \frac{\partial \widehat{\sigma}^2}{\partial t} \\
&= \frac{1}{2\widehat{\sigma}^2} \frac{\partial \overline{P'^2}}{\partial t} \\
&= \frac{1}{\widehat{\sigma}^2} \overline{P' \left(\frac{\partial P}{\partial t} - \frac{\partial \overline{P}}{\partial t} \right)} \\
&= \frac{1}{\widehat{\sigma}^2} \overline{P' \left(NC - \boldsymbol{v} \cdot \boldsymbol{\nabla}_\theta P \right)} \\
&= \frac{1}{\widehat{\sigma}^2} \overline{P' \left(NC - (\overline{\boldsymbol{v}} + \boldsymbol{v}') \cdot \boldsymbol{\nabla}_\theta (\overline{P} + P') \right)} \\
&= -\frac{1}{2\widehat{\sigma}^2} \overline{\boldsymbol{v}} \cdot \boldsymbol{\nabla}_\theta \overline{P'^2} + \frac{1}{\widehat{\sigma}^2} \overline{P' \left(NC - \boldsymbol{v}' \cdot \boldsymbol{\nabla}_\theta P \right)} \\
&= -\frac{1}{\widehat{\sigma}} \overline{\boldsymbol{v}} \cdot \boldsymbol{\nabla}_\theta \widehat{\sigma} + \frac{1}{\widehat{\sigma}^2} \overline{P' \left(NC - \boldsymbol{v}' \cdot \boldsymbol{\nabla}_\theta P \right)} \quad .
\end{aligned}
\tag{A1}
$$

395   Appendix B: Further details on the growth–rate plots

Figure 2 and Fig. 3, and animations 1 and 2 in the supplementary material show, for the two cyclogenesis cases, the growth–rate
Eq. (1; left-hand-side) in the ensemble standard deviation of potential vorticity on the 315 K isentrope $\widehat{\sigma}_{PV315}$ following the
ensemble–mean horizontal flow $\overline{\boldsymbol{v}}_{315}$. This is derived from background (first–guess) forecasts within the ECMWF operational
EDA (see Appendix E), initialised at 06 and 18 UTC, and for hourly lead–times 0–12 h. Fields of $PV315$ and zonal and

400   meridional winds at 315 K ($u315$ and $v315$) are interpolated to an "N32" reduced Gaussian grid (with 32 latitudes between

the pole and equator). The spatial derivatives within the advection term in Eq. (1) are calculated using spectral transforms to and from a "T42" spherical harmonic representation (total wavenumbers $\leqslant 42$). Note that N32 is sufficient to avoid aliasing of higher harmonics of the quadratic advection term into the T42 representation. To focus on synoptic spatial scales, $PV315$ and the growth–rate of $\widehat{\sigma}_{PV315}$ are then spatially smoothed by multiplying spectral coefficients (with total wavenumber $n > n_s = $

405   21) by $\{n_s(n_s+1)\}/\{n(n+1)\}$. 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 $p_* < 850\,\text{hPa}$) 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.

410   This is obtained on a higher resolution "O80" octahedral grid to give a good symbolic representation (stippling) of rainfall. To focus on synoptic timescales, a 1–day running mean is applied to all fields before plotting. Note that animation 3 in the supplementary material shows the animation for the entire DJF 2019/20 season.

Figure 4 and Fig. 5 and animations 4 and 5 in the supplementary material show, for the two cyclogenesis cases, the growth–rate in the ensemble standard deviation of geopotential height at 250 hPa $\widehat{\sigma}_{Z250}$ following the ensemble–mean horizontal flow

415    $\overline{\boldsymbol{v}}_{250}$ for the four chosen TIGGE models. Animation 6 in the supplementary material shows the entire DJF 2019/20 season. These are derived from ensemble data obtained from the TIGGE archive for forecasts initialised at 00 and 12 UTC, and for lead–times 0, 6, 12 h. All other aspects are the same as for the plots of the $\widehat{\sigma}_{PV315}$ growth–rate with the exception that the plotted $Z250$ field is not spatially smoothed (it is considered already a synoptic–scale field).

Appendix C: Derivation of the "extended spread–error" equation

420   With reference to the schematic shown in Fig. 6, consider initially a single ensemble forecast, and thus consider an initial time t_j for some fixed value of j. Next consider a perfectly reliable forecast/analysis system. As discussed in section 6, the verifying truth T_j should be statistically indistinguishable from any random sampling of the forecast distribution F_j. This distribution need not be Gaussian or even uni–modal, but let it have mean μ_{Fj} and variance σ_{Fj}^2. The verifying analysis distribution A_j is constructed using observations of the truth (amongst other things), so that its mean and variance will be the conditional random

425   variables $\mu_{Aj}|T_j$ and $\sigma_{Aj}^2|T_j$. Again, for a reliable analysis distribution, the T_j should be statistically indistinguishable from any random sampling of A_j. (Because of the conditionality above, it is actually uncertainties in the analysis distribution which can enable this). Now consider a realistic forecast/analysis system which need not be reliable. Let the "underlying" forecast distribution $\widetilde{F}_j$ have mean $\mu_{\widetilde{F}j}$ and variance $\sigma_{\widetilde{F}j}^2$, and the underlying verifying analysis distribution $\widetilde{A}_j$ have mean $\mu_{\widetilde{A}j}$ and variance $\sigma_{\widetilde{A}j}^2$. Finally, let the estimators of these quantities based on the real ensemble (finite samplings of $\widetilde{F}_j$ and $\widetilde{A}_j$) be

430 $\widehat{\mu}_{\widetilde{F}j}$, $\widehat{\sigma}^2_{\widetilde{F}j}$, $\widehat{\mu}_{\widetilde{A}j}$ and $\widehat{\sigma}^2_{\widetilde{A}j}$. The departure of the ensemble–mean forecast $\widehat{\mu}_{\widetilde{F}j}$ from the ensemble–mean analysis $\widehat{\mu}_{\widetilde{A}j}$ can then be written (by following the other lines in Fig. 6) as:

$$
\begin{aligned}
d_j \;&= (\widehat{\mu}_{\widetilde{F}j} - \widehat{\mu}_{\widetilde{A}j}) \\[4pt]
&= (\widehat{\mu}_{\widetilde{F}j} - \mu_{\widetilde{F}j}) + (\mu_{\widetilde{F}j} - \mu_{Fj}) + (\mu_{Fj} - T_j) + (T_j - \mu_{Aj}) + (\mu_{Aj} - \mu_{\widetilde{A}j}) + (\mu_{\widetilde{A}j} - \widehat{\mu}_{\widetilde{A}j}) \\[4pt]
&= \underbrace{(\mu_{Fj} - T_j)}_{\mathbf{1}} + \underbrace{(\widehat{\mu}_{\widetilde{F}j} - \mu_{\widetilde{F}j})}_{\mathbf{2}} + \underbrace{(T_j - \mu_{Aj})}_{\mathbf{3}} + \underbrace{(\mu_{\widetilde{A}j} - \widehat{\mu}_{\widetilde{A}j})}_{\mathbf{4}} + \underbrace{(\mu_{\widetilde{F}j} - \mu_{Fj})}_{\mathbf{5}} + \underbrace{(\mu_{Aj} - \mu_{\widetilde{A}j})}_{\mathbf{6}} \quad,
\end{aligned}
\tag{C1}
$$

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

The terms **1**–**6** in Eq. (C1) would be difficult to quantify in the operational forecast system because they require knowledge
435 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. Define $\mathbb{E}_j[\cdot]$ to be the expectation for a given j. Since j is fixed, so are the distributions F_j and $\widetilde{F}_j$; the expectation is over all samplings $\widehat{F}_j$ of $\widetilde{F}_j$, all potential T_j and corresponding analysis distributions A_j and $\widetilde{A}_j$, and all samplings $\widehat{A}_j$ of $\widetilde{A}_j$. Taking the expectation of the last line in Eq. (C1) we have:

$$
\mathbb{E}_j[d_j] = \mathbb{E}_j\left[\mathbf{5} + \mathbf{6}\right] = \mathbb{E}_j\left[(\mu_{\widetilde{F}j} - \mu_{Fj}) - (\mu_{\widetilde{A}j} - \mu_{Aj})\right] \equiv \beta_{\widetilde{F}j} - \beta_{\widetilde{A}j} \equiv \beta_j \quad,
\tag{C2}
$$

440 since terms **1**–**4** in Eq. (C1) 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. (C1). This involves squared terms, such as $\mathbb{E}_j[\mathbf{1} \cdot \mathbf{1}]$, and cross terms, such as $2\mathbb{E}_j[\mathbf{1} \cdot \mathbf{6}]$. In general the squared terms might be thought to dominate, leading to (in the same
445 order as the last line in Eq. (C1)):

$$
\begin{aligned}
\mathbb{E}_j[d_j^2] \;&\approx \sigma^2_{Fj} + \frac{1}{m}\sigma^2_{\widetilde{F}j} + \sigma^2_{Aj} + \frac{1}{m}\sigma^2_{\widetilde{A}j} + (\beta_{\widetilde{F}j} - \beta_{\widetilde{A}j})^2 + \mathcal{E}_j \\[6pt]
&= \frac{m+1}{m}(\sigma^2_{\widetilde{F}j} + \sigma^2_{\widetilde{A}j}) + \beta_j^2 + \Big\{ \underbrace{(\sigma^2_{Fj} - \sigma^2_{\widetilde{F}j}) + (\sigma^2_{Aj} - \sigma^2_{\widetilde{A}j})}_{\text{Variance deficit}} + \mathcal{E}_j\Big\} \quad.
\end{aligned}
\tag{C3}
$$

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).

450 From Eq. (C3), the expected squared departure (over an infinite set of initial forecast times t_j) can then be written as

$$
\mathbb{E}[d^2] = \frac{m+1}{m}\mathbb{E}[\sigma^2_{\widetilde{F}} + \sigma^2_{\widetilde{A}}] + \mathbb{E}[\beta]^2 + \mathbb{E}[R] \quad,
\tag{C4}
$$

with expected "residual",

$$\mathbb{E}[R] = \mathbb{E}\left[(\sigma_F^2 - \sigma_{\widetilde{F}}^2) + (\sigma_A^2 - \sigma_{\widetilde{A}}^2)\right] \;+\; \mathbb{E}[\mathcal{E}] \;+\; \sigma_\beta^2 \quad, \tag{C5}$$

where $\sigma_\beta^2 = \mathbb{V}[\beta]$ accounts for the explicit replacement of $\mathbb{E}[\beta^2]$ with $\mathbb{E}[\beta]^2$ in Eq. (C4). It can be seen that all terms that involve

455 variations in bias have been moved into the residual.

For an unbiased estimator of Eq. (C4), 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:

$$\boxed{\frac{n}{n-1}\overline{d^2} = \frac{m+1}{m-1}\overline{\left(\widehat{\sigma}_{\widetilde{F}}^2 + \widehat{\sigma}_{\widetilde{A}}^2\right)} + \frac{n}{n-1}\overline{d}^2 + \overline{R}} \quad, \tag{C6}$$

where $\overline{R}$ can be written as the mean of the residuals R_j:

460 $$R_j \equiv \frac{n}{n-1}\left(d_j^2 - \overline{d}^2\right) - \frac{m+1}{m-1}\left(\widehat{\sigma}_{\widetilde{F}j}^2 + \widehat{\sigma}_{\widetilde{A}j}^2\right) \quad, \tag{C7}$$

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. (C5) will depend on the magnitude of the other terms $\mathbb{E}[\mathcal{E}] + \sigma_\beta^2$ in Eq. (C5). The only potentially non–zero terms in $\mathbb{E}[\mathcal{E}]$ appear to be $2\mathbb{E}_j[\mathbf{1}\cdot\mathbf{6}]$ (if the analysis bias is dependent on the true state), $2\mathbb{E}_j[\mathbf{2}\cdot\mathbf{4}]$ (at short lead–times and for small ensemble sizes if the analysis is dependent on background forecasts) and the expected variance of the analysis bias (a

465 component of $\mathbb{E}_j[\mathbf{6}\cdot\mathbf{6}]$). These terms are difficult to quantify but estimation suggests that they are likely to be negligible (the estimation was based on 10 DJF seasons, a 50–member ensemble size, day–2 temperatures at 500 hPa, and analysis departures from "AMSUA" satellite channel 5 brightness temperatures). Thus the only additional term that may need accounting for is the variance in forecast bias, σ_β^2. In the flow–dependent evaluation discussed here, this is (at least partially) accounted for explicitly (and is non–negligible).

470 **Appendix D: Global plots of the extended spread–error equation**

Figure D1 and Fig. D2 show global plots of the extended spread–error equation, applied to geopotential height at 500 hPa at day–2 in their non–rooted form, for DJF 2020/21 and JJA 2021 seasons, respectively. ECMWF displays smaller root–mean–square errors of the ensemble–mean than UKMO, and smaller analysis uncertainty. In all winter stormtrack regions, ECMWF shows larger spread (and over–spread) than UKMO. In polar regions, UKMO shows larger spread (and over–spread) than

475 ECMWF.
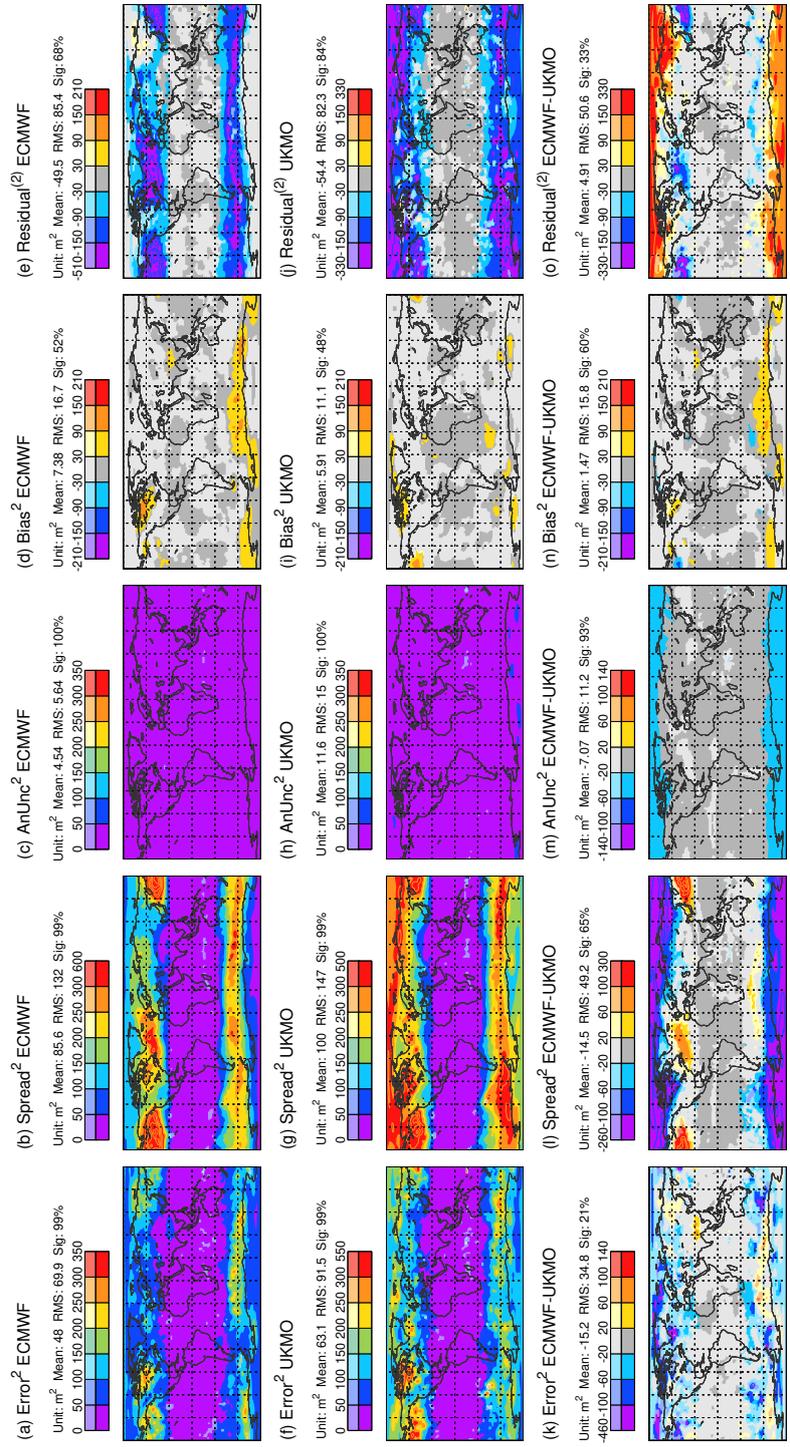
Figure D1. As Fig. 7, but for geopotential height at 500 hPa and showing global maps of the (non–rooted) terms in the extended spread–error budget Eq. (2). Simple differences are shown in the bottom panels.
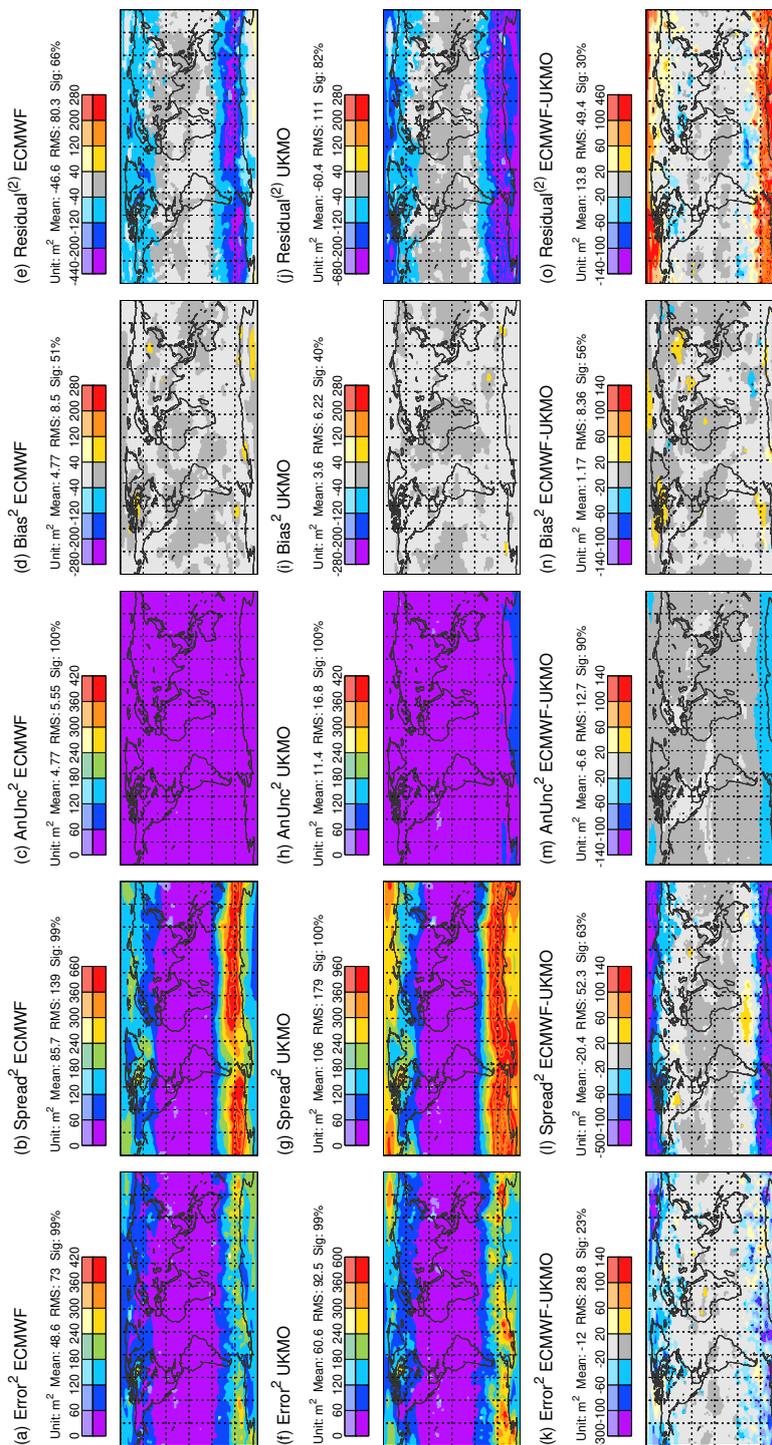
Figure D2. As Fig. D1, but for June – August 2020 Eq. (2).

Appendix E: Salient aspects of ECMWF's Integrated Forecasting System

The sensitivity studies are based on cycle 46r1 of the IFS. This cycle was operational from 11 June 2019 to 20 June 2020. The underlying model 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
480   (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). This model is used within the EDA (Isaksen et al., 2010) and the ENS (Palmer et al., 1992; Molteni et al., 1996). The 50–member EDA has horizontal grid resolution ∼18 km (the model's spectral resolution is T639 and a cubic "octahedral" grid is used so that 4 grid–points represent the smallest waves — the resolution is thus summarised as
485   TCo639) with 137 levels in the vertical (L137) and a 12 min timestep. 4D Variational data assimilation (4DVar, Rabier et al., 2000) is used to extract the information content from many millions of conventional and satellite observations during each 12 h analysis cycle. After initial screening, observations are corrected using Variational Bias Correction (VarBC, Dee, 2004) and randomly perturbed in each EDA member to simulate observation uncertainty (Isaksen et al., 2010). Background uncertainty is dynamically based on the background spread over recent cycles, with a climatological component for stability (Bonavita
490   et al., 2016). 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) analysis. SV perturbations are added to the initial conditions as a pragmatic means of boosting ENS spread over the first 2 days (Molteni and Palmer, 1993; Leutbecher and Lang, 2014). A model uncertainty (MU) parametrization, which partly aims to represent the up–scale response sub–grid–scale uncertainty, is important for the general growth of ENS spread into the medium–range (e.g., to day
495   10). This MU representation is currently based on "Stochastic Perturbation to Physical Tendencies" (SPPT, Buizza et al., 1999), and is applied in the non–linear outer–loops of the EDA as well as in the ENS.

In the experiments here, 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 ∼4 km ENS resolution experiment, the model is run in "single precision" (Váňa et al., 2017; Lang et al., 2021) with resolution
500   TCo2559 L91 and a 4 min timestep.

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, considerable help came from Peter Bechtold, Andrew Dawson, Richard Forbes, Alan Geer, Elias Hólm, Bruce Ingleby, Simon Lang, Inna Polichtchouk, and Gabor Radnoti. 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.

References

510   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.

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.

515   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.

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 ECMWF Ensemble Prediction System,
520   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.: Constraining stochastic parametrisation schemes using high-resolution simulations, Quart. J. Roy. Meteor. Soc., 146, 938–962, https://doi.org/https://doi.org/10.1002/qj.3717, 2020.

525   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,
530   https://doi.org/10.1175/JAS-D-14-0007.1, 2014.

Ehrendorfer, M.: The Liouville Equation and Its Potential Usefulness for the Prediction of Forecast Skill. Part I: Theory, Mon. Wea. Rev., 122, 703 – 713, https://doi.org/10.1175/1520-0493(1994)122<0703:TLEAIP>2.0.CO;2, 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,
535   Quart. J. Roy. Meteor. Soc., 144, 1191–1217, https://doi.org/https://doi.org/10.1002/qj.3202, 2018.

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,
540   https://doi.org/https://doi.org/10.1002/qj.3353, 2018.

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.

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.

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.

Humpherys, J., Redd, P., and West, J.: A Fresh Look at the Kalman Filter, SIAM Review, 54, 801–823, https://doi.org/10.1137/100799666, 2012.

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.

Janjić, T., Bormann, N., Bocquet, M., Carton, J. A., Cohn, S. E., Dance, S. L., Losa, S. N., Nichols, N. K., Potthast, R., Waller, J. A., and Weston, P.: On the representation error in data assimilation, Quart. J. Roy. Meteor. Soc., 144, 1257–1278, https://doi.org/https://doi.org/10.1002/qj.3130, 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.

Lewis, J. M.: Roots of Ensemble Forecasting, Mon. Wea. Rev., 133, 1865–1885, https://doi.org/10.1175/MWR2949.1, 2005.

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.

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.

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.

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.

Portmann, R., González-Alemán, J. J., Sprenger, M., and Wernli, H.: How an uncertain short-wave perturbation on the North Atlantic wave guide affects the forecast of an intense Mediterranean cyclone (Medicane Zorbas), Weather Clim. Dynam., 1, 597–615, https://doi.org/10.5194/wcd-1-597-2020, 2020.

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.

585 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
590 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.

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.

595 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.

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,
600 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,
605 1989.

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.

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.

610 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.