How consistently do ensemble prediction systems represent the growth of atmospheric uncertainty?

Received: 16 June 2025 | Revised: 17 December 2025 | Accepted: 21 February 2026

DOI: 10.1002/qj.70168

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

How consistently do ensemble prediction systems represent the growth of atmospheric uncertainty?

Douglas Wood[1] | **Suzanne Gray**[1] | **John Methven**[1] | **Mark Rodwell**[2]

[1]University of Reading, Reading, UK

[2]European Centre for Medium Range Weather Forecasts, Reading, UK

Correspondence
Douglas Wood, University of Reading, Reading, UK.
Email: doug.wood@pgr.reading.ac.uk

Abstract

Recent studies have used ensemble-spread-based diagnostics to understand atmospheric sources of uncertainty, and its growth and propagation. We aim to provide confidence in the diagnostics-based approach, and understand its limitations, by comparing Eulerian and Lagrangian geopotential height uncertainty growth diagnostics calculated from 12 different ensemble prediction systems available in The International Grand Global Ensemble. In addition to a Lagrangian growth-rate diagnostic that uses the ensemble mean velocity (LGR), an alternative form (LGR2) is derived that includes all the uncertainty transport terms and that, for variables with tracer-like characteristics, also represents the non-conservative sources of uncertainty. Good agreement between ensemble prediction systems is found in the magnitude and distribution of all growth-rate diagnostics applied in the midlatitude and polar regions for a lead-time range starting at 48 hr and ending at between 96 hr and 192 hr, with the best agreement for the largest ensemble size of 50 members. In these situations, ensemble-spread-based diagnostics provide a consistent measure of uncertainty growth. However, at shorter and longer lead times, and in tropical regions, ensemble prediction system dependence is greater. In addition, LGR and LGR2 are compared for the fields of geopotential height and potential vorticity (a tracer-like variable) using data from the European Centre for Medium-Range Weather Forecasts operational archive. For geopotential height, LGR2 is found to include a significant non-advective component; therefore, restricting the use of this diagnostic to variables that have tracer-like characteristics is recommended. This study provides confidence in, and constraints on, the use of spread-based diagnostics for understanding the sources and transport of uncertainty growth that limit predictability.

KEYWORDS

atmospheric uncertainty, ensemble prediction systems, predictability, uncertainty growth

1 | INTRODUCTION

The atmosphere exhibits chaotic behaviour resulting from the nonlinear dynamics that govern its development over time (Lorenz, 1963). This chaotic behaviour leads to an inherent uncertainty in the weather forecasting process and imposes limits on predictability. Ensemble forecasting, since the (UK) Met Office quasi-operational probabilistic ensemble forecast system was developed in 1985 (Palmer, 2019), has provided a method for quantifying this uncertainty and attributing conditional probabilities to future events. In ensemble forecasting uncertainties in the initial conditions, the boundary conditions and the modelling of physical processes are assessed by running a forecast model with small perturbations relative to a control forecast (Lewis, 2005). The spread of the ensemble members provides information on the forecast uncertainty. For a given ensemble prediction system, the growth of this spread varies both spatially and with time through the forecast, indicating that forecast uncertainty is dependent on the conditions; that is, it is flow dependent (Rodwell *et al.*, 2018). Synoptic-scale variations in the growth of the spread can, therefore, provide insight into the weather patterns and physical processes that are influencing the atmosphere's sensitivity to small perturbations, as in Baumgart and Riemer (2019). Here, we investigate whether forecast uncertainty growth, represented by the evolution of ensemble-spread-based diagnostics, does provide a useful measure of the atmosphere's sensitivity to small perturbations, or whether it is significantly affected by other factors associated with the ensemble prediction system configuration, including its numerical weather prediction (NWP) model and perturbation scheme. The aim of this work is to provide confidence in the use of ensemble-spread-based diagnostics for understanding the physical processes that are associated with the growth and transport of atmospheric uncertainty, where atmospheric uncertainty is defined here as the sensitivity of the future state of the atmosphere to small changes in the current conditions. To achieve this aim, spread-based diagnostics obtained from a number of different operational ensemble prediction systems are compared.

The scale interactions that occur within the atmosphere result in a limit to predictability, termed intrinsic predictability, even if the initial and boundary conditions are known almost perfectly and a nearly perfect NWP model is used (Lorenz, 1969). In practice, uncertainties in the initial and boundary conditions, combined with the capability of current NWP models, place a practical predictability limit on the ensemble prediction system forecasts. The modelling goal in NWP is to provide reliable ensemble forecasts at all lead times. This goal requires ensemble prediction systems to represent, as well as possible, the processes and dynamics that govern intrinsic predictability. Therefore, in principle, ensemble prediction systems should produce similar estimates for uncertainty (in terms of ensemble spread), how it grows and is transported from one location to another and how this evolution is related to flow features, such as cyclones. Note that other potential sources of similarity in the uncertainty estimated by two ensemble prediction systems include any common features in the formulation of their underlying NWP models and the common set of observations that are available for use in their respective data assimilation systems. In the Baumgart and Riemer (2019) study, data from a single ensemble prediction system, the European Centre for Medium-Range Weather Forecasts (ECMWF) ensemble, was used to investigate the extratropical transition of Tropical Storm *Karl*. They found that a maximum in the ensemble variance of potential vorticity (PV), implying larger than average forecast uncertainty, occurred where ex-Tropical Storm *Karl* interacted with the large PV gradient in the midlatitudes. The growth of PV variance in this region was attributed to uncertainties associated with the moist processes involved in the development of an extratropical cyclone. However, their study did not investigate whether this PV relationship was also obtained using data from other ensemble prediction systems. A comparison between the uncertainty growth rate obtained from four different ensemble prediction systems by Rodwell and Wernli (2023) showed large differences in the uncertainty growth rate, calculated from the 250 hPa geopotential height ensemble variance, in a region of the North Atlantic over the first 24 hr of an example forecast. Their results imply that, for short lead times, ensemble-prediction-system-dependent factors make a significant contribution to forecast uncertainty growth.

Leutbecher and Palmer (2008) separate forecast errors into initial condition errors and model errors, although they also note that these are not independent as an older model run is used as the background forecast in the data assimilation process that generates the initial conditions for subsequent model runs. In addition, boundary condition errors will also contribute to the overall forecast error. Information on the contribution of these sources of error to the forecast uncertainty is provided by the spread–error relationship used to assess the performance of an ensemble prediction system. This relationship compares the forecast spread measured by the ensemble variance of a given variable with the error in the ensemble mean forecast, the latter calculated as the squared difference between the ensemble mean and the value obtained from the verifying analysis. Over multiple forecasts, the ensemble mean error would be expected to equal the ensemble spread if the ensemble prediction system including its associated NWP model was either (a) perfect—that

is, it had an infinite number of members and there were no ensemble prediction system errors—or (b) account had been taken of the errors associated with using a finite number of members and an imperfect ensemble prediction system by, for example, adjusting the magnitude of the initial perturbations to obtain the required spread–error statistics within a given lead-time range. However, for a real ensemble prediction system, a difference between the ensemble mean error and ensemble spread is expected at all, or most, lead times. The development of this difference with forecast lead time was investigated by Park *et al.* (2008) for eight ensemble prediction systems using data obtained from The International Grand Global Ensemble (TIGGE) archive (Swinbank *et al.*, 2016). In their study, the ensemble standard deviation and the root-mean-squared error of the 500 hPa geopotential height, when averaged across the Northern Hemisphere, were typically within ±10 m during the first 4 days of the forecast. Sources of the difference between the spread and error (Buizza *et al.*, 2005) are associated with NWP-related factors and how they affect each perturbed member, including:

- the design and resolution of the NWP model grid in both the horizontal and vertical directions;
- simplifications to the dynamic and thermodynamic equations and the numerical methods for solving these equations; and
- the methods used for parametrisation of subgrid-scale processes.

Sources of difference are also associated with the ensemble prediction system design, including:

- the finite number of ensemble members;
- the methods for generating the initial perturbations; and
- the stochastic perturbation methods used to represent the effects of subgrid-scale model uncertainty.

Ensemble-spread diagnostics are only useful as a measure of the atmospheric uncertainty if these sources of difference do not contribute a significant proportion of the forecast uncertainty. Furthermore, if their contribution is small across a number of ensemble prediction systems, the forecast spread for these ensembles should be similar. Thus, by comparing ensemble-spread-based diagnostics across ensemble prediction systems, the quality of these diagnostics as a measure of atmospheric uncertainty can be assessed.

In this study, spread-based diagnostics are compared across 12 ensemble prediction systems to understand whether the differences in the uncertainty growth rates obtained by Rodwell and Wernli (2023) are typical, or whether there are circumstances under which spread-based diagnostics are relatively independent of the ensemble prediction system. The intent of the work is to determine the limitations of ensemble-spread-based diagnostics for understanding both the sources of atmospheric uncertainty and how this atmospheric uncertainty is transported. Understanding these limits will support their use in investigating the atmospheric conditions and processes that contribute to uncertainty. The benefit of the approach, when compared with the use of forecasting skill measures obtained directly from the spread–error relationship, is that it enables the spatial distribution of forecast uncertainty and uncertainty growth for each ensemble prediction system to be compared for individual forecasts.

In Section 2, the uncertainty and uncertainty growth diagnostics are introduced. The ensemble prediction systems, selection of the diagnostic variables, and comparison methods used within this study are addressed in Section 3. The results of the comparisons and their interpretation are presented in Section 4. Finally, Section 5 summarises the conclusions and discusses the implications of the findings.

2 | UNCERTAINTY DIAGNOSTICS

Forecast uncertainty can be quantified using either the ensemble variance (Baumgart & Riemer, 2019),

$$\mathrm{Var}_X = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2, \tag{1}$$

or, as selected for this current study, the ensemble standard deviation (Rodwell & Wernli, 2023),

$$\sigma_X = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}, \tag{2}$$

where X is the variable of interest, N the ensemble size, and $\overline{X}$ the ensemble mean $\left(\frac{1}{N}\sum_{i=1}^{N}X_i\right)$. In addition to σ_X, which provides useful information on the spatial distribution of forecast uncertainty at a given forecast lead time, measures of uncertainty growth can also be used to understand the sources and transport of uncertainty. The Eulerian growth is a standard measure of local instantaneous growth and can be expressed in a growth tendency,

$$\mathrm{EGT}_X = \frac{\partial \sigma_X}{\partial t}, \tag{3}$$

or growth rate form in which the growth tendency has been normalised:

$$\mathrm{EGR}_X = \frac{1}{\sigma_X}\frac{\partial \sigma_X}{\partial t}. \tag{4}$$

EGT_X is independent of spread for cases in which the local uncertainty grows linearly, whereas EGR_X is independent of spread for cases in which the local uncertainty grows exponentially.

An alternative growth-rate diagnostic with a Lagrangian form has been used by Rodwell and Wernli (2023). In this Lagrangian growth diagnostic, the uncertainty growth following the ensemble mean velocity $\overline{\mathbf{v}}$ is calculated. The diagnostic can be expressed in either a growth tendency form,

$$\mathrm{LGT}_X = \frac{D_{\overline{\mathbf{v}}}\sigma_X}{Dt} = \frac{\partial \sigma_X}{\partial t} + \overline{\mathbf{v}} \cdot \nabla \sigma_X, \tag{5}$$

or in a growth rate form,

$$\mathrm{LGR}_X = \frac{1}{\sigma_X}\frac{D_{\overline{\mathbf{v}}}\sigma_X}{Dt} = \frac{1}{\sigma_X}\left(\frac{\partial \sigma_X}{\partial t} + \overline{\mathbf{v}} \cdot \nabla \sigma_X\right). \tag{6}$$

The advecting velocity assumption in Equations (5) and (6) can be investigated by deriving the equation for LGR_X using an approach based on Baumgart and Riemer (2019). For each ensemble member i, the transport equation for the variable X_i can be written as

$$\frac{\partial X_i}{\partial t} + \mathbf{v}_i \cdot \nabla X_i + \mathrm{NonAdv}_i = S_i, \tag{7}$$

where $\mathbf{v}_i$ is the wind velocity, NonAdv_i represents any non-advective terms in the conservative dynamics, and S_i represents non-conservative source and sink terms. For tracer-like variables, such as PV on a constant potential temperature (θ) surface, $\mathrm{NonAdv}_i = 0$. Expressing X_i, $\mathbf{v}_i$, NonAdv_i, and S_i as a sum of the ensemble mean values and a perturbation from the mean $\overline{X} + \delta X_i$, $\overline{\mathbf{v}} + \delta \mathbf{v}_i$, $\overline{\mathrm{NonAdv}} + \delta\mathrm{NonAdv}_i$ and $\overline{S} + \delta S_i$ respectively, multiplying throughout by $2\delta X_i/(n-1)$, rewriting $2\delta X_i(\partial \delta X_i/\partial t)$ as $\partial \delta X_i^2/\partial t$, and summing over all ensemble members leads to the following equation:

$$\begin{aligned}
&\frac{2}{n-1}\sum_{i=1}^{n}\delta X_i \frac{\partial \overline{X}}{\partial t} + \frac{1}{n-1}\sum_{i=1}^{n}\frac{\partial \delta X_i^2}{\partial t} \\
&\quad + \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i(\overline{\mathbf{v}} + \delta \mathbf{v}_i)\cdot\nabla(\overline{X} + \delta X_i) + \\
&\quad \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i(\overline{\mathrm{NonAdv}} + \delta\mathrm{NonAdv}_i) \\
&= \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i(\overline{S} + \delta S_i).
\end{aligned} \tag{8}$$

The first term in Equation (8) and those involving $\overline{\mathrm{NonAdv}}$ and $\overline{S}$ sum to zero. The second term is the time derivative of the ensemble variance, $\partial \sigma_X^2/\partial t$. The third term can be separated into

$$\begin{aligned}
&\frac{2}{n-1}\left(\sum_{i=1}^{n}\delta X_i \overline{\mathbf{v}}\cdot\nabla\overline{X} + \sum_{i=1}^{n}\delta X_i\overline{\mathbf{v}}\cdot\nabla\delta X_i \right. \\
&\left. \quad + \sum_{i=1}^{n}\delta X_i \delta\mathbf{v}_i\cdot\nabla(\overline{X} + \delta X_i)\right).
\end{aligned} \tag{9}$$

The first term in Equation (9) also sums to zero, and the second term can be rewritten as

$$\frac{2}{n-1}\sum_{i=1}^{n}\delta X_i\overline{\mathbf{v}}\cdot\nabla\delta X_i = \overline{\mathbf{v}}\cdot\frac{\sum_{i=1}^{n}\nabla\delta X_i^2}{(n-1)} = \overline{\mathbf{v}}\cdot\nabla\sigma_X^2. \tag{10}$$

Thus, Equation (8) becomes

$$\begin{aligned}
&\frac{\partial \sigma_X^2}{\partial t} + \overline{\mathbf{v}}\cdot\nabla\sigma_X^2 + \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i\delta\mathbf{v}_i\cdot\nabla(\overline{X} + \delta X_i) \\
&\quad + \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i\delta\mathrm{NonAdv}_i \\
&= \frac{2}{n-1}\sum_{i=1}^{n}\delta X_i\delta S_i,
\end{aligned} \tag{11}$$

or, dividing by $2\sigma_X^2$ to rewrite Equation (11) in terms of the growth rate of σ_X,

$$\begin{aligned}
&\frac{1}{\sigma_X}\frac{\partial \sigma_X}{\partial t} + \frac{1}{\sigma_X}\overline{\mathbf{v}}\cdot\nabla\sigma_X + \frac{1}{(n-1)\sigma_X^2}\sum_{i=1}^{n}\delta X_i\delta\mathbf{v}_i\cdot\nabla X_i \\
&\quad + \frac{1}{(n-1)\sigma_X^2}\sum_{i=1}^{n}\delta X_i\delta\mathrm{NonAdv}_i \\
&= \frac{1}{(n-1)\sigma_X^2}\sum_{i=1}^{n}\delta X_i\delta S_i.
\end{aligned} \tag{12}$$

Moving the non-advective term to the right-hand side of Equation (12), so that the left-hand side only contains terms that can be directly calculated from the values of $\mathbf{v}_i$ and X_i, and rewriting it in terms of either LGR_X or EGR_X,

$$\begin{aligned}
\mathrm{LGR2}_X &= \mathrm{LGR}_X + D_X \\
&= \mathrm{EGR}_X + A_X + D_X \\
&= \mathrm{NC}_X - \mathrm{NA}_X,
\end{aligned} \tag{13}$$

where NC_X and NA_X are used to represent the sums that include the non-conservative and non-advective terms, A_X is advection of the uncertainty by the ensemble mean wind, and D_X is an additional transport term that is generated when summing the left-hand side of Equation (7) over the ensemble members:

$$D_X = \frac{1}{(n-1)\sigma_X^2}\sum_{i=1}^{n}\delta X_i\delta\mathbf{v}_i\cdot\nabla X_i. \tag{14}$$

The notation $\mathrm{LGR2}_X$ has been introduced as the sum of LGR_X, the Lagrangian growth rate based on advection by the ensemble mean, and D_X, the additional transport term. A growth tendency diagnostic, $\mathrm{LGT2}_X$, can be obtained by multiplying $\mathrm{LGR2}_X$ by σ_X.

The uncertainty growth-rate equation, Equation (13), and its individual terms have the same form as those for the evolution of tracer variance in turbulent flow; for example, in eq. 13.67 of Vallis (2017). Although the perturbations in the tracer variance equation are due to the eddy component of the flow, whereas in Equation (13) they are the differences between values for the individual members and the ensemble mean, the comparison between these two equations provides an insight into the role of D_X. In the Vallis (2017) tracer variance equation, the equivalent term is diffusive since the tracer flux often satisfies the downgradient flux hypothesis, $\delta X_i \delta \mathbf{v}_i \approx -\kappa \nabla \overline{X}$, where κ can be interpreted as the eddy diffusivity. Here, D_X acts to diffuse perturbations in the tracer across the ensemble mean tracer gradient. D_X would then act to increase ensemble spread (like a plume in turbulent dispersion) as well as to smooth the ensemble mean gradient. In this regime of behaviour $\mathrm{LGR2}_X$ describes "transport", including the action of advection and diffusion associated with the perturbation velocities between ensemble members. However, it must be established by calculation whether the output of ensemble prediction systems is consistent with this behaviour. Rodwell and Wernli (2023) moved the D_X term onto the right-hand side of Equation (13) in their appendix B since they found that typically $D_X < 0$ and therefore contributes to the increase of ensemble variance with time following the mean flow ($\mathrm{LGR}_X > 0$) (note that in their derivation they used D to instead represent the diabatic and frictional terms that are included here within the non-conservative term NC_X). An attribute of $\mathrm{LGR2}_X$ is that it isolates the effects of non-conservative, and other non-advective, processes on the rate of increase of ensemble spread. For tracer-like variables $\mathrm{NA}_X = 0$, $\mathrm{LGR2}_X = \mathrm{NC}_X$, where NC_X is the covariance between the variable and the source terms. Thus, as an example, if θ on a constant PV surface were chosen as the tracer-like variable, $\mathrm{LGR2}_\theta$ would be the covariance between θ and the heating term $\dot{\theta}$, and a positive value of $\mathrm{LGR2}_\theta$ would be consistent with a growth in the uncertainty of θ.

For a perfect ensemble prediction system, the $\mathrm{LGR2}_X$ value in Equation (13) equals the sum of the non-conservative and non-advective terms associated with atmospheric uncertainty. However, for real ensemble systems, a component of these non-conservative and non-advective terms is associated with ensemble prediction system errors. Therefore, when comparing the $\mathrm{LGR2}_X$ diagnostic calculated across several ensemble prediction systems, any differences in the $\mathrm{LGR2}_X$ values are also a relative measure of these errors. In addition to an understanding of these relative errors, a comparison of each of the three growth diagnostics (EGR_X, LGR_X and $\mathrm{LGR2}_X$) across several different ensemble prediction systems provides information on the relative magnitude of the A_X and D_X terms.

Although these comparisons are not sufficient to eliminate the possibility that all the ensemble prediction systems are subject to the same systematic error, the independence of growth-rate diagnostics from the amplitude of the spread reduces the potential for any systematic errors arising from the use of the same set of observations. In addition, previous work investigating the spread–error relationship for different ensemble prediction systems, including Park *et al.* (2008) and Rodwell *et al.* (2018), has demonstrated good agreement between ensemble spread and error when measured over a large number of forecasts. This agreement suggests that the ensemble prediction systems are not subject to a large systematic error. Thus, consistency in diagnostics provides confidence that they are representative of atmospheric uncertainty and are not strongly influenced by the design and simplifications associated with the ensemble prediction systems.

3 | ENSEMBLE PREDICTION SYSTEMS AND METHODS OF ANALYSIS

An overview of the ensemble prediction systems and the sources of data used for the comparisons are described in Section 3.1. Section 3.2 discusses the selection of variables for calculating the diagnostics, and Section 3.3 discusses the methods used to compare the diagnostics calculated for the different ensemble prediction systems.

3.1 | Ensemble prediction systems

Data stored within the TIGGE archive (Swinbank *et al.*, 2016) was used in this study. Global data from the 13 ensemble prediction systems that have participated in TIGGE since 2006 have been stored within this archive. As organisations have joined and left the collaboration since its inception, a maximum of 12 have participated at any point in time (since July 2020). This study has used data for the 12 ensemble prediction systems that were participating in 2022. Table 1 lists these ensemble prediction systems and some of their key attributes according to information documented within TIGGE. Model grid and resolution, initial perturbation strategy and ensemble size vary across the ensemble prediction systems, thereby reducing the possibility that where consistent results are

TABLE 1 Ensemble prediction systems used for comparison analysis.

Model code	Forecasting centre	Model	Grid/resolution	Initial perturbation strategy[a]	N[b]	T_{max}[c] (hr)
BoM	Australian Bureau of Meteorology	GE3	Arakawa C (N400), 0.45° lon × 0.3° lat	ETKF	18	240
CMA	China Meteorological Administration	GRAPES GFS	Arakawa C, 0.5° lon × 0.5° lat	SV	31	360
DWD	Deutscher Wetterdienst	ICON 2.6.0	R2B06, 40 km	LETKF	40	180
ECCC	Environment and Climate Change Canada	GEPS 6.0.0	Yin-Yang, 39 km	EnKF	21	384
ECMWF	European Centre for Medium-Range Weather Forecasts	IFS 47r3	TCo639	SV, EDA	51	360
IMD	India Meteorological Department	GEFS	T1534	ETR, EnKF	21	240
JMA	Japan Meteorological Agency	GEPS 2103	T479	SV, LETKF	51	264
KMA[d]	Korea Meteorological Administration	EPSG 1306	Arakawa C (N320), 0.5625° lon × 0.375° lat cubed sphere grid about 32 km	ETKF	25	288
		EPSG 2201		LETKF	26	288
MF	MeteoFrance	PEAR P4	TL798	SV	35	48[e]
NCEP	US National Centres for Environmental Prediction	GEFS v12	C384	EnKF	31	384
NCMRWF	National Centre for Medium Range Weather Forecasting (India)	NEPS v2	Arakawa C (N1024), 0.175° lon × 0.117° lat	ETKF	12	240
UKMO	UK Met Office	MOGREPS-G v16	Arakawa C (N640) 0.28125° lon × 0.1875° lat	ETKF	18	174

[a]EDA: ensemble of data assimilations; EnKF: ensemble Kalman filter; ETKF: ensemble transform Kalman filter; ETR: ensemble transform with rescaling; LETKF: local ensemble transform Kalman filter; SV: singular vectors.
[b]Number of perturbed members, including, where applicable, the control forecast.
[c]Maximum duration of the 0000 UTC and 1200 UTC forecasts.
[d]KMA introduced new model in July 2022.
[e]MF model duration is longer for 0600 UTC and 1800 UTC forecasts. However, 0000 UTC and 1200 UTC forecasts were used for consistency.

obtained this is attributable to common features of the ensemble prediction system design. The duration of the 0000 UTC and 1200 UTC forecasts stored in TIGGE are also included in Table 1, with some centres running forecasts at additional start times. The results presented in this article are based on the 0000 UTC forecasts. For some systems, interpolation is used to translate the model output to their specific storage grid within the TIGGE archive. Further interpolation is also employed, in combination with the selection of the download grid, when retrieving the data. These interpolation processes may contribute to differences in the diagnostics calculated for the different ensemble prediction systems.

3.2 | Selection of variables

The preferred method for the ensemble prediction system comparison is to calculate diagnostics for a tracer-like variable, such as PV on a constant θ surface, or θ on a constant PV surface. However, insufficient data are available within the TIGGE archive to calculate diagnostics for these variables; therefore, comparisons were limited to the use of five variables, all of which are available on several pressure levels: geopotential height Z, specific humidity q, temperature T, and the zonal and meridional wind components u and v respectively. These variables are archived every 6 hr throughout the duration of each forecast. For the Rodwell

and Wernli (2023) study, which also compared ensemble prediction systems using TIGGE data, geopotential height on the 250 hPa surface Z_{250} was used. The horizontal distribution of Z can be considered to be a smoothed version of the PV field (Hoskins *et al.*, 1985) and, therefore, the evolution of the synoptic-scale structure of this Z field should be similar to the evolution of the equivalent structure in the PV field. Thus, Z_{250} has also been selected as the main variable for this study. The 250 hPa pressure surface was selected as it is in the upper troposphere or lower stratosphere spanning the tropopause, an important region in the development of weather systems in the midlatitudes. Selecting the 250 hPa pressure surface also enables the results to be compared with those of Rodwell and Wernli (2023).

Although data are not available within the TIGGE archive to enable a comparison of diagnostics between ensemble systems for PV, approximate PV values can be calculated for the ECMWF ensemble prediction system using data from the ECMWF operational archive. This archive stores PV data on the $\theta = 320$ K surface but without any corresponding wind velocity data. However, approximate PV values can be calculated for a constant-pressure surface using the vertical component of relative vorticity ξ_p and static stability in pressure coordinates:

$$\mathrm{PV} \approx -g(\xi_p + f)\frac{\partial \theta}{\partial p}, \tag{15}$$

where f is the Coriolis parameter. This approximate form of PV enables diagnostics for Z_{250} and PV_{250} to be compared with the vertical derivative of θ estimated from the centred finite difference of data on the 200 hPa and 300 hPa pressure surfaces.

3.3 | Comparison methods

To enable comparisons between ensemble prediction systems to be consistent, data were retrieved from TIGGE using a $1° \times 1°$ latitude–longitude grid. In addition to the standard latitude–longitude grid, data were also retrieved using a rotated "latitude–longitude" grid with the grid "South Pole" located at 180° W, 0° N and grid "North Pole" located at 0° W, 0° N. The rotated grid was used to avoid both large variations in resolution and to maintain a more regular grid lattice near the Earth's poles when calculating diagnostics in the Arctic and Antarctic regions. Spatial derivatives were calculated from the gridded data using finite differences. Diagnostics were calculated at six-hourly intervals throughout the duration of forecasts initiated at 0000 UTC on the first day of each month in 2022, with the number of forecasts limited by the storage capacity and computational expense. These 12 forecasts were selected to understand whether the comparisons showed any obvious seasonal effects.

Diagnostics were calculated at the midpoint between consecutive forecast lead times (i.e., at 0300 UTC, 0900 UTC, 1500 UTC, and 2100 UTC on each day of each forecast) using the following centred averages for the σ_x values, centred differences for the time derivatives, and spatial derivatives calculated from the centred average σ_x values:

$$\sigma_{i,j}^{n+\frac{1}{2}} = \frac{\sigma_{i,j}^{n+1} + \sigma_{i,j}^{n}}{2}, \tag{16}$$

$$\left(\frac{\delta\sigma}{\delta t}\right)_{i,j}^{n+\frac{1}{2}} = \frac{\sigma_{i,j}^{n+1} - \sigma_{i,j}^{n}}{\Delta t}, \tag{17}$$

$$\left(\frac{\delta\sigma}{\delta x}\right)_{i,j}^{n+\frac{1}{2}} = \frac{\sigma_{i+1,j}^{n+\frac{1}{2}} - \sigma_{i-1,j}^{n+\frac{1}{2}}}{2\Delta x}, \tag{18}$$

$$\left(\frac{\delta\sigma}{\delta y}\right)_{i,j}^{n+\frac{1}{2}} = \frac{\sigma_{i,j+1}^{n+\frac{1}{2}} - \sigma_{i,j-1}^{n+\frac{1}{2}}}{2\Delta y}, \tag{19}$$

where, here, n is the time index and i and j are the indices in the x-direction and y-direction of the grid. The calculation of derivatives using centred differences and centred averages will introduce a numerical error into the calculation of diagnostics. However, the magnitude of this numerical error should be similar for cases in which the diagnostic values obtained from different ensemble prediction systems are also similar. In addition to the individual ensemble prediction systems, the diagnostics were also calculated for a combined ensemble. To ensure that the same number of members was used for all forecast lead times, the MeteoFrance ensemble members (for which only 48-hr forecasts were available) were not included in the 319-member combined ensemble and diagnostics were only calculated for lead times between 9 hr and 171 hr.

Qualitatively, the spatial distribution of the diagnostic values (e.g., maps of LGR) between ensemble prediction systems can be compared visually. Quantitative comparisons were obtained using the area-weighted spatial averages of each diagnostic, and the Pearson correlation coefficient r_{AB} for the spatial correlation between maps of the same diagnostics for ensemble prediction systems A and B. The area-weighted spatial averages enable a generic comparison of the uncertainty and uncertainty growth behaviour between ensemble prediction systems but do not provide information on whether they consistently predict high and low uncertainty growth values for the same synoptic-scale features. This limitation is addressed by the

spatial correlation method that compares the spatial distribution of the diagnostic values. Area-weighted averages were calculated directly from the $1° \times 1°$ grid. However, prior to calculating the spatial correlation coefficient between the diagnostic maps for each ensemble prediction system pair, spatial smoothing to synoptic scales was applied to the final diagnostic values using a Gaussian kernel filter with the standard deviation of the kernel filter set to one grid space; that is, 1° on the standard or rotated grid. The r_{AB} values for each lead time were averaged over all pairs of ensemble prediction systems to provide a combined spatial correlation value $\overline{r_{AB}}$. The quantitative measures were calculated globally and, to understand whether the results were latitude dependent, for 30° bands of latitude. Data for calculating the diagnostic values were retrieved on the standard grid for the tropical regions (30° S to 0° S and 0° N to 30° N), the rotated grid for the polar regions (90° S to 60° S and 60° N to 90° N), and both grids for midlatitude regions (60° S to 30° S and 30° N to 60° N).

4 | EVOLUTION OF FORECAST UNCERTAINTY

Section 4.1 discusses the evolution of the forecast uncertainty diagnostics calculated using Z_{250} for the midlatitude and polar regions. Section 4.2 documents the growth behaviour in the tropical regions. Section 4.3 discusses the impact of ensemble size on the spatial correlation between ensemble prediction systems for the $\mathrm{LGR}_{Z_{250}}$ diagnostic. Finally, Section 4.4 compares uncertainty growth-rate diagnostics calculated from the ECMWF operational archive using Z_{250} and the approximate value of PV_{250} as the variable.

4.1 | Extratropical regions

Figure 1 shows, for five of the diagnostic values, how the quantitative comparison between the 12 ensemble prediction systems evolves with lead time during the 0000 UTC May 1, 2022, forecast. Figure 1a,d,g,j,m shows the area-weighted average values of the diagnostics in the Northern Hemisphere midlatitudes, and Figure 1b,e,h,k,n shows the area-weighted average values of the same diagnostics in the Arctic. Figure 1c,f,i,l,o shows the average spatial correlation coefficient $\overline{r_{AB}}$, the spatial distribution of diagnostic values for each ensemble pair averaged across all ensemble prediction system pairs, for both these regions over the first 168 hr of the same forecast. The error bars shown represent the standard deviation of the r_{AB} values. The area-weighted averages for the $\mathrm{EGT}_{Z_{250}}$ and $\mathrm{LGT2}_{Z_{250}}$ diagnostics (not shown) are similar to those for $\mathrm{LGT}_{Z_{250}}$, and the $\overline{r_{AB}}$ for $\mathrm{EGT}_{Z_{250}}$ and $\mathrm{LGT2}_{Z_{250}}$ (not shown) are almost identical to those for $\mathrm{EGR}_{Z_{250}}$ and $\mathrm{LGR2}_{Z_{250}}$ respectively. Figure 1 shows that the area-weighted average values for $\mathrm{EGR}_{Z_{250}}$, $\mathrm{LGR}_{Z_{250}}$, and $\mathrm{LGR2}_{Z_{250}}$ evolve in almost the same way, implying that the advection term $A_{Z_{250}}$ and the diffusion term $D_{Z_{250}}$ are small when spatially averaged over a large surface area. The evolution of the area-weighted averages for the various diagnostics can be divided into three time periods. The first period covers the initial 48 hr of the forecast, whereas the second and third periods cover the longer lead times. The transition between the second and third periods is more gradual than between the first and second and takes place at lead times between 96 hr and 192 hr.

In the northern midlatitude region the initial value of $\overline{\sigma_{Z_{250}}}$, a measure of the spread of a given ensemble's initial perturbations, ranges by a factor of more than five from 1.8 m (National Centre for Medium Range Weather Forecasting) to 9.5 m (Deutscher Wetterdienst) (Figure 1a). By 48 hr, $\overline{\sigma_{Z_{250}}}$ has increased to between 10.8 m (UK Met Office) and 20.2 m (Deutscher Wetterdienst). Thus, although $\overline{\sigma_{Z_{250}}}$ has grown for all ensembles over these first 48 hr of the forecast, the range between the lowest and highest values has only increased from 7.7 m to 9.4 m, indicating that differences in growth are not a strong function of the initial spread during this period. This is the period in which the spatially averaged growth rates ($\overline{\mathrm{EGR}_{Z_{250}}}$, $\overline{\mathrm{LGR}_{Z_{250}}}$, and $\overline{\mathrm{LGR2}_{Z_{250}}}$) are at a maximum and is characterised by inconsistencies in the values of these diagnostics between the ensemble systems. Although differences are also found for the average growth tendency values, they are much more consistent than the average growth-rate values. The evolution of $\overline{LGT_{Z_{250}}}$ is shown as an example in Figure 1g. The results indicate that some ensemble prediction systems have growth tendency values that are at the higher or lower end of the range throughout the first 144 hr of the forecast. The better consistency in the average growth tendencies implies that the differences in the average growth rates are due to the relative magnitude of the $\sigma_{Z_{250}}$ values amplifying the differences in the growth tendency diagnostics. The second period is characterised by almost constant $\overline{\mathrm{LGR}_{Z_{250}}}$ and $\overline{\mathrm{LGR2}_{Z_{250}}}$ values, implying that $\overline{\sigma_{Z_{250}}}$ grows approximately exponentially in the northern midlatitudes, and there is much more consistency between the average growth-rate values calculated for each of the ensemble prediction systems than at shorter lead times. In this midlatitude region, the transition to the third period is associated with a peak in the average growth tendency values. During this third period, both average growth rates and growth tendencies start to decrease towards zero and $\overline{\sigma_{Z_{250}}}$ values tend towards a maximum, indicating that the spread is no longer growing

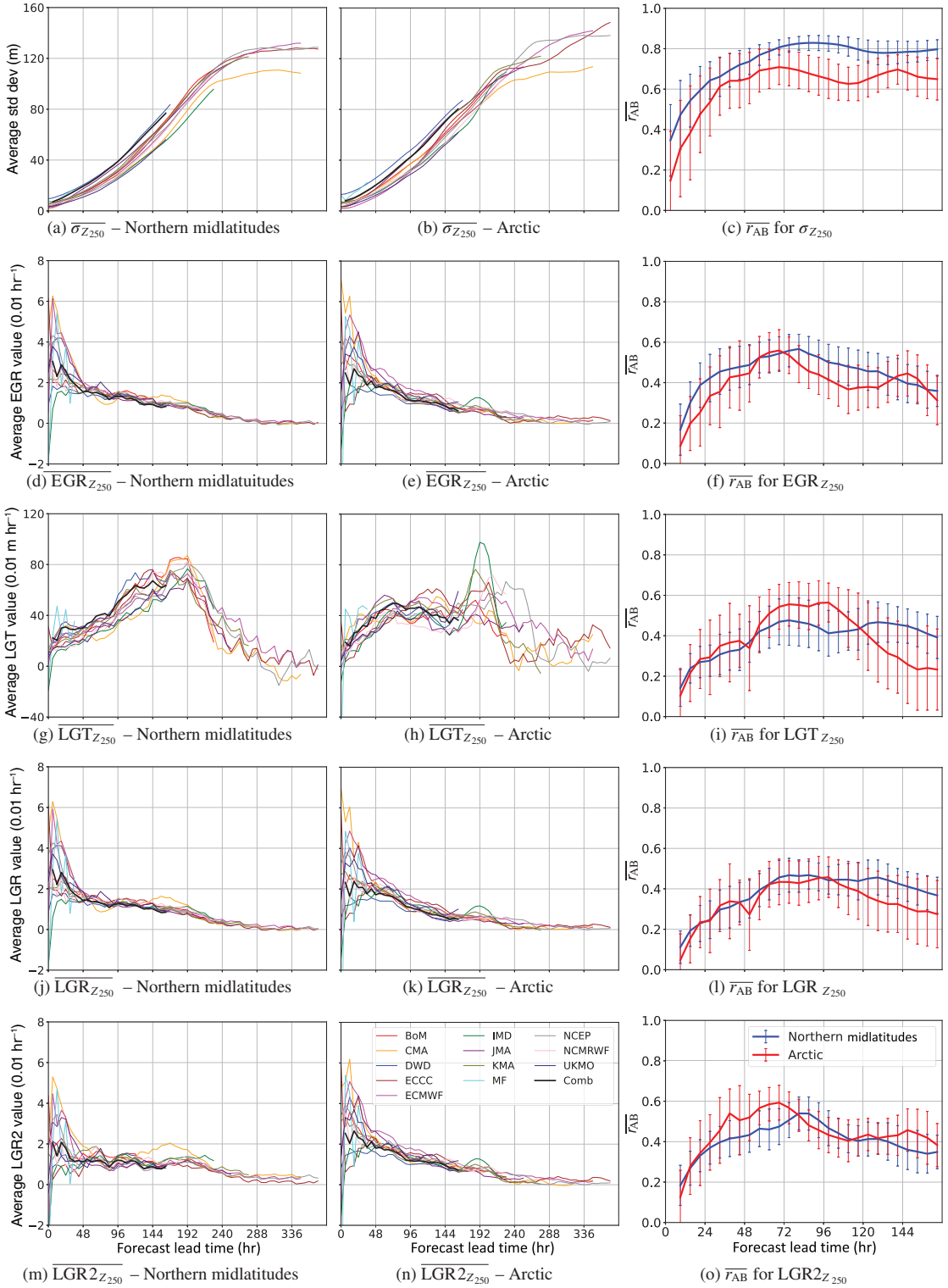
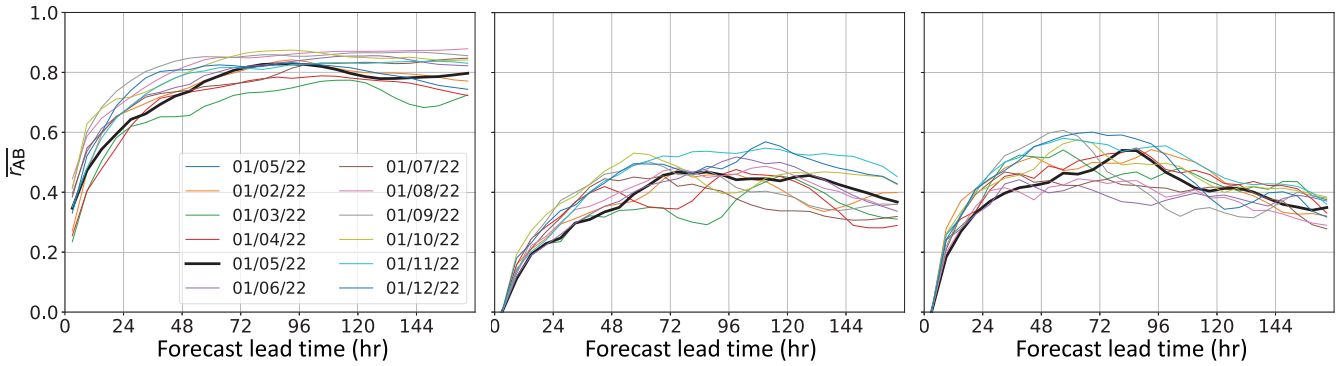
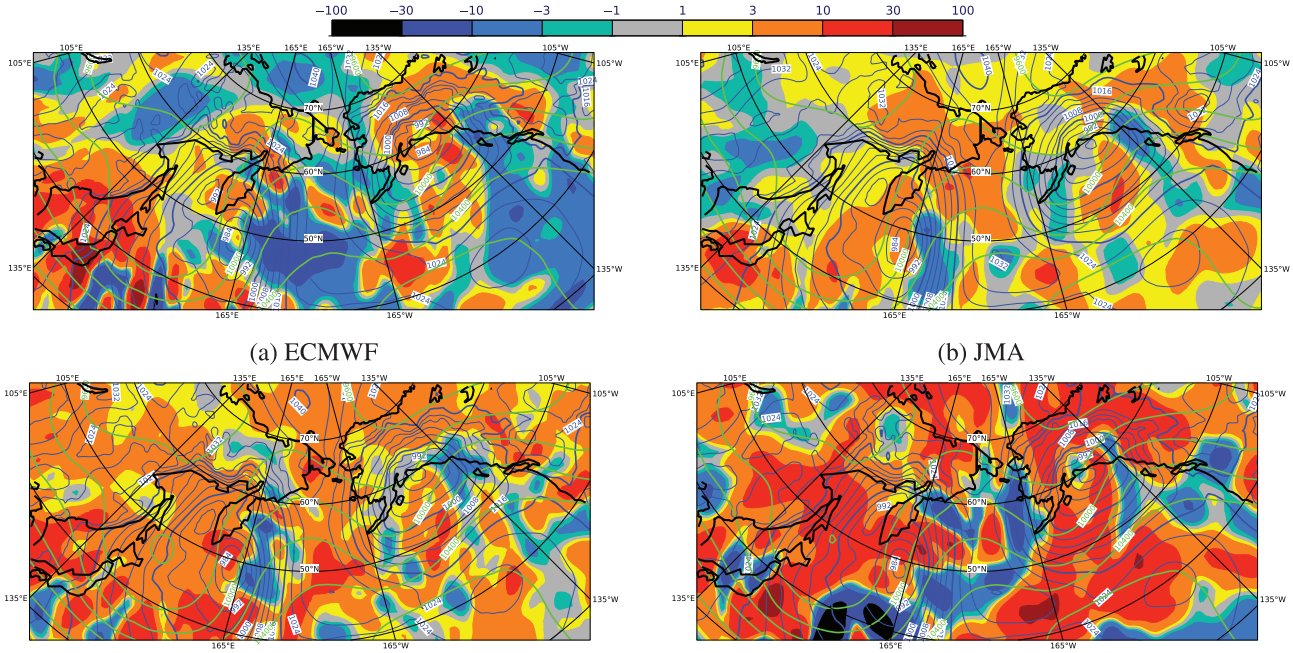


FIGURE 1 Quantitative comparison of diagnostics between the 12 ensemble prediction systems across the northern midlatitudes (between 30° N and 60° N) and Arctic (between 60° N and 90° N) as a function of lead time for forecasts starting at 0000 UTC May 1, 2022. Left and middle panels show the area-weighted average values for five of the diagnostics in the two regions and the right panels show the corresponding average correlation coefficients ($\overline{r_{AB}}$) for both regions. Error bars represent the standard deviation of the correlation coefficient for individual ensemble prediction system pairs. BoM: Australian Bureau of Meteorology; CMA: China Meteorological Administration; DWD: Deutscher Wetterdienst; ECCC: Environment and Climate Change Canada; ECMWF: European Centre for Medium-Range Weather Forecasts; IMD: India Meteorological Department; JMA: Japan Meteorological Agency; KMA: Korea Meteorological Administration; MF: MeteoFrance; NCEP: National Centres for Environmental Prediction; NCMRWF: National Centre for Medium Range Weather Forecasting (India); UKMO: UK Met Office. [Colour figure can be viewed at wileyonlinelibrary.com]



(a) $\overline{r_{AB}}$ for $\sigma_{Z_{250}}$ (b) $\overline{r_{AB}}$ for $\mathrm{LGR}_{Z_{250}}$ (c) $\overline{r_{AB}}$ for $\mathrm{LGR2}_{Z_{250}}$

FIGURE 2 Comparison of average correlation coefficient $\overline{r_{AB}}$ for the northern midlatitudes (between 30° N and 60° N) for the 12 forecasts analysed. The forecast start times were 0000 UTC on the first day of each month in 2022 with the 0000 UTC May 1, 2022, highlighted in black. [Colour figure can be viewed at wileyonlinelibrary.com]



(a) ECMWF (b) JMA

(c) NCEP (d) CMA

FIGURE 3 Smoothed $\mathrm{LGR}_{Z_{250}}$ values (10^{-2} hr^{-1}) at 3 hr into the forecast initiated at 0000 UTC January 1, 2022, for four ensemble prediction systems, in a region covering part of the North Pacific and Arctic: (a) European Centre for Medium-Range Weather Forecasts (ECMWF); (b) Japan Meteorological Agency (JMA); (c) National Centres for Environmental Prediction (NCEP); (d) China Meteorological Administration (CMA). Also shown are the mean-sea-level pressure from the respective control forecasts (blue contours) and the ensemble mean value of Z_{250} (green contours). [Colour figure can be viewed at wileyonlinelibrary.com]

exponentially and becomes saturated at a value determined by the ensemble prediction system's climatological range.

In the Arctic region, growth-rate values are also consistent across the ensemble prediction systems after a lead time of 48 hr. However, in contrast to the northern midlatitude region, the second period is characterised by almost constant $\overline{\mathrm{LGT}_{Z_{250}}}$ values, indicating growth is approximately linear. Although growth rates are observed to decrease after the first 48-hr period, the transition between the second and third periods can still be defined by the decrease in the growth tendency values that follows the period of linear growth.

Low $\overline{r_{AB}}$ values, the spatial correlation coefficient calculated for each ensemble pair averaged over all ensemble pairs, are obtained at the start of the forecast for both the

northern midlatitude and Arctic regions, with $\overline{r_{AB}}$ typically between 0.1 and 0.2 at a lead time of 9 hr for all diagnostics. $\overline{r_{AB}}$ values then increase throughout the remainder of this first period. During the second period, $\overline{r_{AB}}$ values are relatively constant. This is the lead-time range in which the maximum values are obtained before values start to decrease at the start of the third period. Comparing the diagnostics at a given lead time, $\sigma_{Z_{250}}$ has the highest $\overline{r_{AB}}$ values, with maxima of 0.70 and 0.85 for the northern midlatitude and Arctic regions respectively. $\mathrm{LGR}_{Z_{250}}$ has the lowest $\overline{r_{AB}}$ values, with maxima for both regions of approximately 0.45. These lower $\overline{r_{AB}}$ values for the growth rates compared with those for $\sigma_{Z_{250}}$ are to be expected, as derivative fields tend to be noisier, with this additional noise typically resulting in a reduction in spatial correlation. The similar evolution of the diagnostics for the Arctic and northern midlatitude regions shown in Figure 1 is typical of both the Northern and Southern Hemisphere polar (60° N to 90° N and 60° S to 90° S) and midlatitude (30° N to 60° N and 30° S to 60° S) regions for all 12 forecasts investigated. The variation in $\overline{r_{AB}}$ values across the forecasts is shown on Figure 2 for three of the diagnostics in the northern midlatitudes, with the results for the 0000 UTC May 1 forecast highlighted in black. This representative forecast is used to illustrate the quantitative results throughout the remainder of the article.

An example qualitative comparison of $\mathrm{LGR}_{Z_{250}}$ between ensemble prediction systems is shown in Figures 3–5. This example, from the 0000 UTC January 1 forecast, has been selected as it shows some large-scale mid-latitude synoptic features. Figure 3 shows the spatial distribution of the diagnostic in a region of the North Pacific and Arctic for four of the ensemble prediction systems (ECMWF, Japan Meteorological Agency [JMA], National Centres for Environmental Prediction, and China Meteorological Administration) at a lead time of 3 hr. Figures 4 and 5 are the spatial distributions for the same ensemble prediction systems and diagnostic at lead times of 51 hr and 159 hr respectively. These three lead times correspond to the early part of each of the periods discussed earlier herein. In addition to the diagnostic values, contours of both the mean-sea-level pressure from the control forecast and of the ensemble mean geopotential height are also shown on each of these figures. The following remarks are not specific to the four ensemble prediction systems included in these figures and would be typical of any combination of the 12 ensemble prediction systems (not shown). The main synoptic-scale features shown in Figure 3 are a trough to the south of Alaska with a midlatitude storm system (145° W, 59° N) that has a central pressure below 984 hrPa, a ridge south of the Bering Strait, and a second trough and low-pressure system south of the Kamchatka Peninsula (152° E, 53° N). The large differences between the spatial distribution of $\mathrm{LGR}_{Z_{250}}$ values calculated for the four ensemble prediction systems are in line with the results of Rodwell and

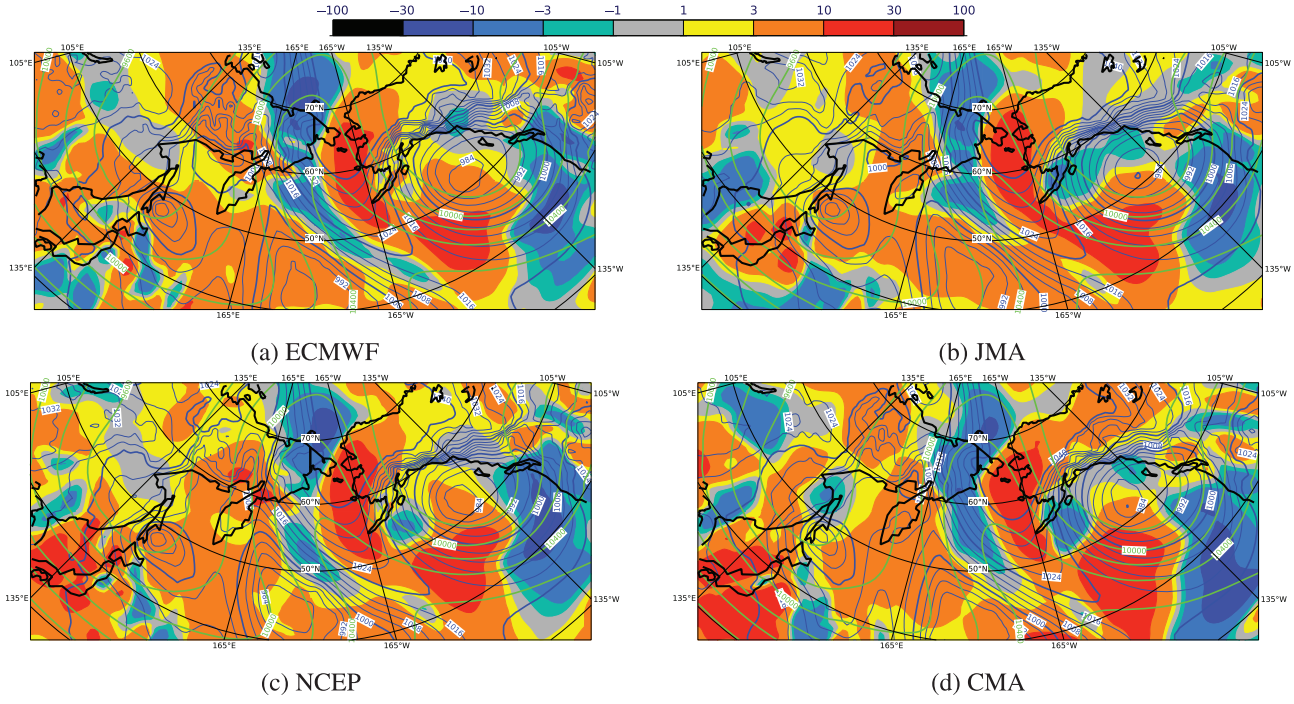


FIGURE 4 As Figure 3 but for a lead time of 51 hr. [Colour figure can be viewed at wileyonlinelibrary.com]

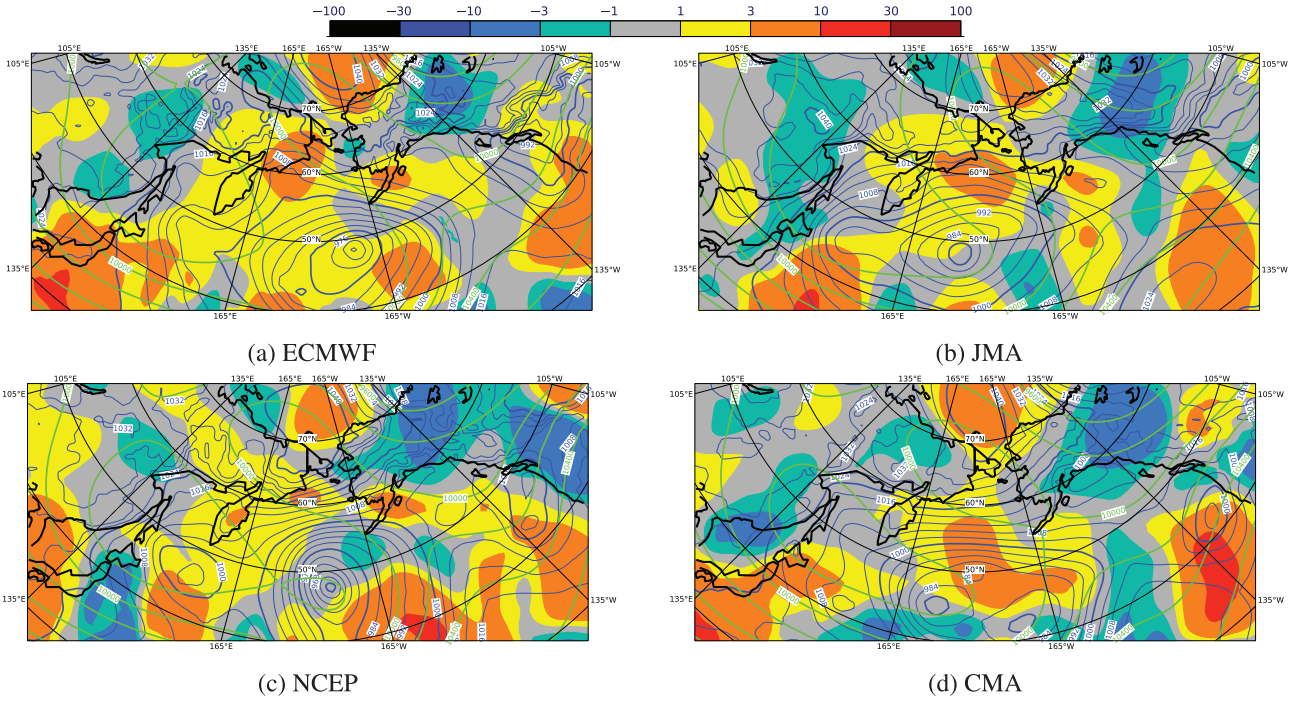


FIGURE 5 As Figure 3 but for a lead time of 159 hr. [Colour figure can be viewed at wileyonlinelibrary.com]

Wernli (2023). The large qualitative differences are also consistent with the low $\overline{r_{AB}}$ values at early lead times shown in Figure 1l. Zonally, the trough and ridge features identified by the mean geopotential height contours in Figure 3 remain relatively static over the following 48 hr period, but Figure 4 shows that their meridional extent has increased as an Ω-block develops with the ridge axis at approximately 180° W. The main low-pressure system to the south of Alaska has moved to the southeast and maintained its strength (140° W, 52° N), whereas the system to the south of Kamchatka has moved north and decayed. A new low-pressure system (145° E, 48° N) has developed to the north of Japan. At a lead time of 51 hr, the spatial distribution of $LGR_{Z_{250}}$ is similar for the four ensemble prediction systems, which is consistent with the higher values of $\overline{r_{AB}}$ obtained during this second period. By 159 hr into the forecast (Figure 5), the ridge in the Z_{250} contours has rotated to become northwest–southeast, the original low-pressure region has dissipated, and the main synoptic feature in the central North Pacific is the low-pressure region (170° W, 49° N) to the west of the ridge that had been developing to the north of Japan at a lead time of 51 hr. Although the four ensemble systems show the same main synoptic features, there are differences in both their location and intensity. These differences in location and intensity are consistent with the study by Froude (2010) that tracked the location of extratropical cyclones in the control and ensemble mean forecast using TIGGE data. By this lead time there is an overall decrease in the magnitude of the $LGR_{Z_{250}}$ values for all ensemble prediction systems, and the spatial differences in this diagnostic between systems has increased. These results are consistent with the decrease in $\overline{LGR_{Z_{250}}}$ and the $\overline{r_{AB}}$ values for $LGR_{Z_{250}}$ in the third period at the longer lead (Figure 1j–l).

To understand the relationship between synoptic systems and the spatial distribution of the other six diagnostics, Figure 6 shows the spatial distribution of these diagnostics obtained from the ECMWF ensemble prediction system for the same 51 hr lead time and 0000 UTC January 1, 2022, forecast as Figure 4a. The $EGR_{Z_{250}}$ and $EGT_{Z_{250}}$ values (Figure 6c,d) show the same main features, as they only differ by the variable scaling of $\sigma_{Z_{250}}$. This similarity is also the case for the $LGR_{Z_{250}}$ and $LGT_{Z_{250}}$ values (Figures 4a and 6b) and the $LGR2_{Z_{250}}$ and $LGT2_{Z_{250}}$ values (Figure 6e,f). Figure 6a shows that the largest uncertainty $\sigma_{Z_{250}}$ is to the south of the main low-pressure region and to the west of the ridge. The $EGR_{Z_{250}}$ and $EGT_{Z_{250}}$ values (Figure 6c,d) show that the local growth tendency and growth rate are also large to the south of the main low-pressure area. However, local growth rates within the ridge and upstream trough are more complex. There is strong growth within the ridge and some small areas of decay in regions where large $\sigma_{Z_{250}}$ values are observed. By comparing Figures 6c and 4a or Figure 6b,d, the importance of the advection term $A_{Z_{250}}$ in the $LGR_{Z_{250}}$ and $LGT_{Z_{250}}$ diagnostics can be understood. Including the advection term based on the mean ensemble velocity

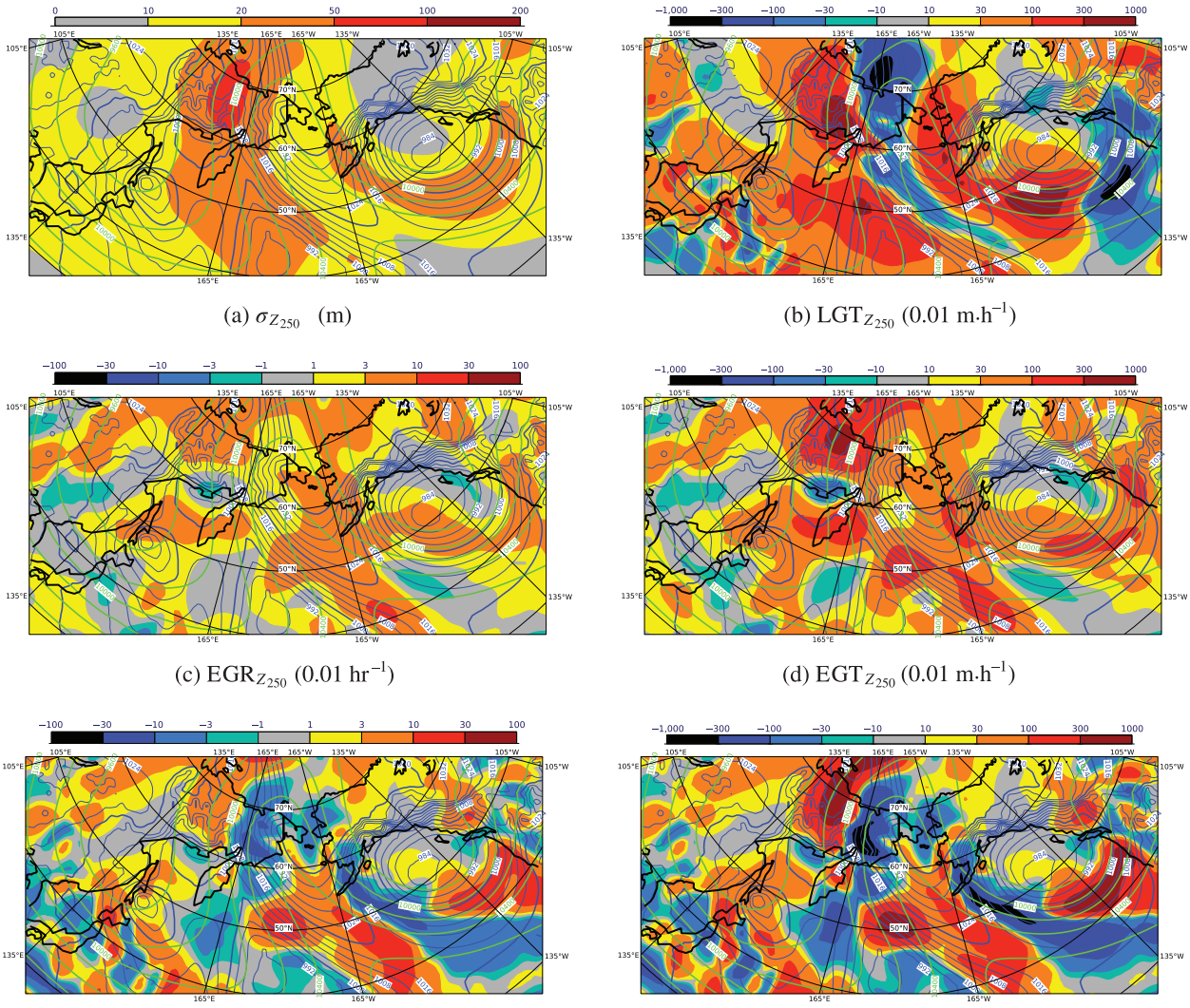


FIGURE 6 Smoothed diagnostic values for the European Centre for Medium-Range Weather Forecasts ensemble prediction system in a region covering part of the North Pacific and Arctic, at 51 hr into the forecast initiated at 0000 UTC January 1, 2022. Also shown are the mean-sea-level pressure from the respective control forecasts (blue contours) and the ensemble mean value of Z_{250} (green contours). [Colour figure can be viewed at wileyonlinelibrary.com]

introduces large areas with negative growth to the east of the main low-pressure region and within the ridge structure. In addition, the differences in the spatial distribution of $\mathrm{LGR}_{Z_{250}}$ and $\mathrm{LGR2}_{Z_{250}}$ (or $\mathrm{LGT}_{Z_{250}}$ and $\mathrm{LGT2}_{Z_{250}}$) highlight the importance of the diffusion term $D_{Z_{250}}$. In many locations, a positive $\mathrm{LGR}_{Z_{250}}$ corresponds to a negative $\mathrm{LGR2}_{Z_{250}}$ value, and vice versa.

The evolution of the area-weighted averages and spatial correlation values over time shown in Figure 1 are attributed to three main components of forecast uncertainty growth:

1. The initial ensemble perturbation structure and spread.
2. The underlying chaotic nature of the atmosphere.
3. The differences, errors, or simplifications within the ensemble prediction system's representation of the atmosphere and atmospheric processes.

The behaviour of the diagnostics in the first 48 hr of the forecast can be attributed to the variation in the structure and spread of the initial perturbations for the different ensemble prediction systems. These variations are likely to be associated with the differences in the perturbation schemes used and their initial growth characteristics. A comparison of the spread–error relationship evolution between ensemble prediction systems was investigated by Park *et al.* (2008) and showed a minimum in the difference between the spread and root-mean-squared error

values typically occurred at lead times of between 2 and 5 days. Thus, ensemble prediction systems with perturbation schemes that exhibit larger initial uncertainty would typically have smaller initial growth rates. However, as ensemble prediction systems have similar values for each of the diagnostics after the first 48 hr of the forecast, the effect of the differences in the initial spread characteristics appears to reduce as lead time increases and perturbations align with the main features of the synoptic-scale flow. By the start of the second period, the effect of differences in the spread of the initial perturbations no longer appears to be the dominant factor in the uncertainty growth rate, growth-rate diagnostics are relatively ensemble prediction system independent and regions of underlying large atmospheric uncertainty growth are being consistently identified. However, at these lead times, the differences in the methods used for modelling and the parametrisation of atmospheric processes will start to lead to some divergence in the location and intensity of weather systems. This limits the maximum $\overline{r_{AB}}$ values and results in a decline in the spatial correlation after the peak that occurs in the second period. These differences in location and intensity of synoptic systems increase through the third period. In addition, Figure 1a indicates that the lead time at which saturation occurs and the maximum spread values are both ensemble system dependent, a result that is attributed to differences in the underlying model design. Combined with the continued divergence in the forecast location and intensity of weather systems, the saturation differences underpin the steady decrease in $\overline{r_{AB}}$ values through the third period.

4.2 | Tropical regions

Figure 7 shows the evolution of selected diagnostics in the northern Tropics for the 12 ensemble prediction systems during the 0000 UTC May 1, 2022, forecast. The results shown are for Z_{250}, enabling a direct comparison with the results for the extratropics. As the main source of variability in the Tropics is associated with equatorial waves (Judt, 2020), geopotential height is a useful variable

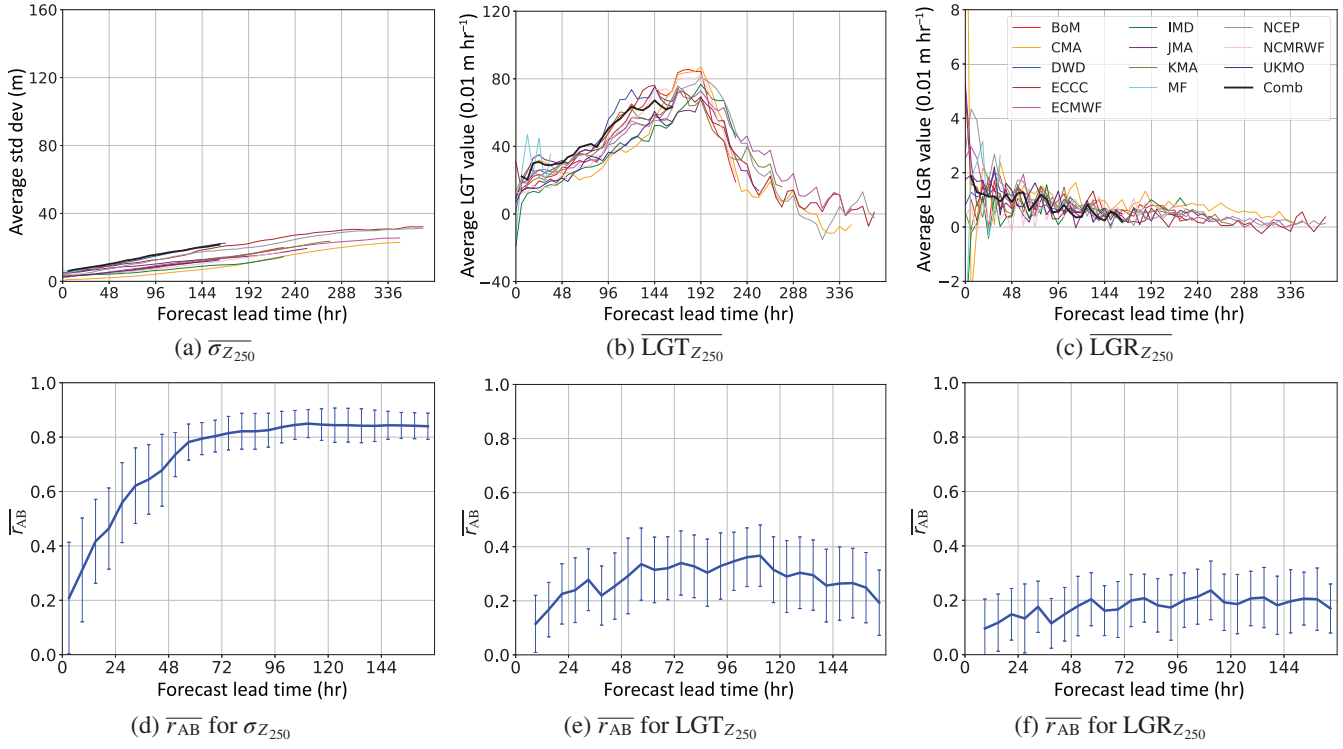


FIGURE 7 Quantitative comparison of diagnostics between the 12 ensemble models across the northern Tropics (between 0° N and 30° N) as a function of lead time for forecasts starting at 0000 UTC May 1, 2022: (a)–(c) area-weighted average values for $\sigma_{Z_{250}}$, $\mathrm{LGT}_{Z_{250}}$, and $\mathrm{LGR}_{Z_{250}}$; (d)–(f) average correlation coefficients $\overline{r_{AB}}$ for the same diagnostics. Error bars represent the standard deviation of the correlation coefficient for individual ensemble prediction system pairs. BoM: Australian Bureau of Meteorology; CMA: China Meteorological Administration; DWD: Deutscher Wetterdienst; ECCC: Environment and Climate Change Canada; ECMWF: European Centre for Medium-Range Weather Forecasts; IMD: India Meteorological Department; JMA: Japan Meteorological Agency; KMA: Korea Meteorological Administration; MF: MeteoFrance; NCEP: National Centres for Environmental Prediction; NCMRWF: National Centre for Medium Range Weather Forecasting (India); UKMO: UK Met Office. [Colour figure can be viewed at wileyonlinelibrary.com]

for understanding uncertainty in this region. Figure 7a–c shows the area-weighted average values for the $\sigma_{Z_{250}}$, $\mathrm{LGT}_{Z_{250}}$, and $\mathrm{LGR}_{Z_{250}}$ diagnostics. Figure 7d–f shows the average spatial correlation coefficient $\overline{r_{\mathrm{AB}}}$, the spatial distribution of diagnostic values for each ensemble pair averaged across all ensemble prediction system pairs, for the same three diagnostics. The evolution of the $\mathrm{EGT}_{Z_{250}}$ and $\mathrm{LGT2}_{Z_{250}}$ area-weighted average values and spatial correlation coefficients (not shown) are similar to those for the $\mathrm{LGT}_{Z_{250}}$ diagnostic. Likewise, the evolution of these values for $\mathrm{EGR}_{Z_{250}}$ and $\mathrm{LGR2}_{Z_{250}}$ (not shown) are similar to those for $\mathrm{LGR}_{Z_{250}}$. The main features of these plots are typical of the evolution of these diagnostics in both the northern and southern Tropics and for all 12 forecast start times. $\overline{\sigma_{Z_{250}}}$ increases approximately linearly up to a lead time of at least 240 hr, and the growth tendency $\overline{\mathrm{LGT}_{Z_{250}}}$ is almost constant throughout this period. The growth rates in the region are much smaller than those in the midlatitudes and polar regions (Figure 1). In addition, short time-scale fluctuation in $\overline{\mathrm{LGR}_{Z_{250}}}$ occur throughout the forecast. The average spatial correlation between ensemble prediction system pairs at the start of the forecast is poor, and the $\overline{r_{\mathrm{AB}}}$ values for $\mathrm{LGR}_{Z_{250}}$ of between 0.1 and 0.2 are similar to those in the extratropical regions. Although $\overline{r_{\mathrm{AB}}}$ for $\sigma_{Z_{250}}$ increases continuously over the first 96 hr, at which time values above 0.8 indicate the ensemble prediction systems have similar spatial distributions of ensemble spread, $\overline{r_{\mathrm{AB}}}$ values for the growth diagnostics, $\mathrm{LGT}_{Z_{250}}$ and $\mathrm{LGR}_{Z_{250}}$, remain low throughout the forecast, peaking at only 0.2. Judt (2020) also found that the Tropics had lower medium- to long-term growth rates than the extratropics and improved predictability at longer lead times. There may be some localised circumstances, such as the development of tropical cyclones, when growth rates are significant. However, these were not obvious on the smoothed $1° \times 1°$ grid data for the forecasts investigated (not shown). In addition to Z_{250}, uncertainty growth associated with several velocity-based variables was also investigated in the Tropics including the zonal velocity u_{250}, meridional velocity v_{250}, and kinetic energy $\frac{1}{2}(u_{250}^2 + v_{250}^2)$. The uncertainty growth-rate results for these velocity-based variables (not shown) were similar to those for Z_{250}, with all three uncertainty growth rates (EGR, LGR, and LGR2) that were much smaller than in the extratropical regions, were linear, and that also included diurnal fluctuations. The spatial correlation coefficients for all variables, at almost all lead times, were smaller in the northern Tropics than in the northern midlatitudes, although the differences were less for the velocity-based variables than for the Z_{250} values shown in Figures 1 and 7.

The short time-scale fluctuations in $\overline{\mathrm{LGR}_{Z_{250}}}$ values are attributed to the diurnal cycle associated with tropical convective processes combined with the uneven zonal distribution of land masses and, therefore, the main regions of convection. Figure 8 shows the $\mathrm{LGR}_{Z_{250}}$ development across the tropical region during day 3 of the ECMWF forecast initiated at 0000 UTC May 1, 2022. High $\mathrm{LGR}_{Z_{250}}$ values are obtained over southeast Asia (region centred on 110° E, 10° N) at 0300 UTC, central Africa (region centred on 30° E, 5° N) at 0900 UTC, western Africa (region centred on 0° E, 10° N) and South America (region centred on 60° W, 5° S) at 1500 UTC. The high values in these highlighted locations are followed by negative $\mathrm{LGR}_{Z_{250}}$ values approximately 6 hr later. This suggests that uncertainty growth over each land area is largest when convection initiates in the morning and then starts to decay once convection rates have peaked. This daily cycle is repeated through the other days of the forecast. Although not investigated here, tropical weather patterns, including monsoons and the Madden–Julian oscillation, may also contribute to an uneven zonal distribution of the uncertainty growth rate. The poorer spatial correlation between the ensemble system growth-rate diagnostics than for the extratropics, combined with the dominance of diurnal convection relative to the development of weather systems, may limit the usefulness of these diagnostics for understanding the sources, growth, and transport of atmospheric uncertainty in tropical regions.

4.3 | Effect of ensemble size

Figure 9a shows the relationship between the r_{AB} values of the $\mathrm{LGR}_{Z_{250}}$ diagnostic plotted against the average ensemble size for each ensemble pair. The values are for lead times of between 48 hr and 72 hr from the 0000 UTC May 1, 2022, forecast. Although there is significant scatter, the figure shows a trend towards higher r_{AB} values with increasing ensemble size. This trend is observed for lead times of more than 48 hr in most of the extratropical and polar cases. Leutbecher and Palmer (2008) show that a spread correction factor, $(N+1)/(N-1)$, needs to be applied when analysing spread–error relationships to account for the finite ensemble size, implying that, when comparing the spread diagnostics for two perfect ensembles with N_1 and N_2 members, a difference would expected when $N_1 \neq N_2$. However, this is not the source of the relationship between r_{AB} and ensemble size found here, as the exponential form of the $\mathrm{LGR}_{Z_{250}}$ diagnostic in Equation (6) means that growth rates are independent of the absolute value of the forecast spread.

To better understand the relationship between spatial correlation and ensemble size, reduced-member ensembles were generated using the control forecast and a

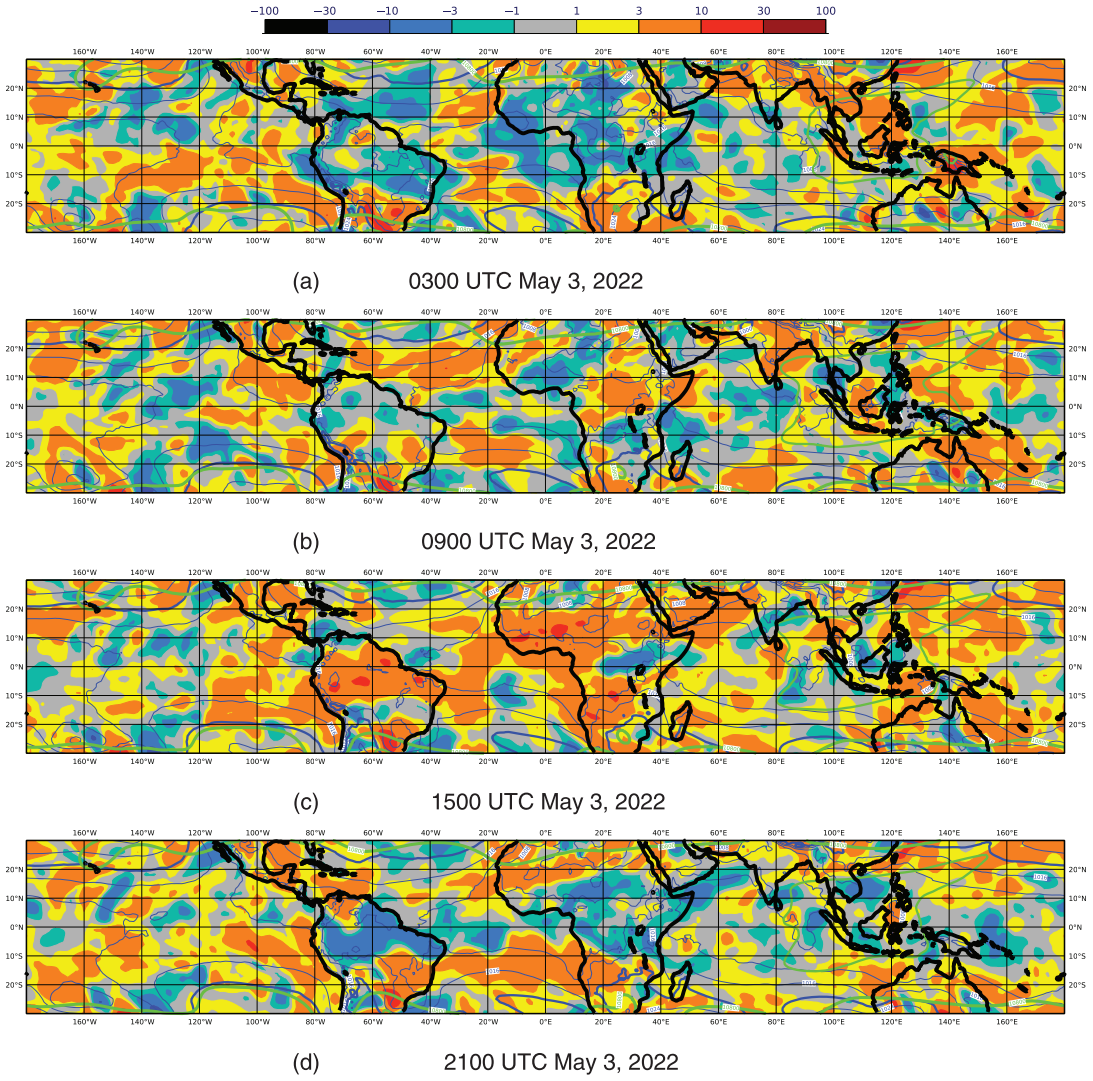


FIGURE 8 Smoothed $\mathrm{LGR}_{Z_{250}}$ values ($10^{-2}\,\mathrm{hr}^{-1}$) in tropical regions for ECMWF forecast initiated at 0000 UTC May 1, 2022. Also shown are the mean-sea-level pressure from the respective control forecast (blue contours) and the ensemble mean value of the Z_{250} (green contours). [Colour figure can be viewed at wileyonlinelibrary.com]

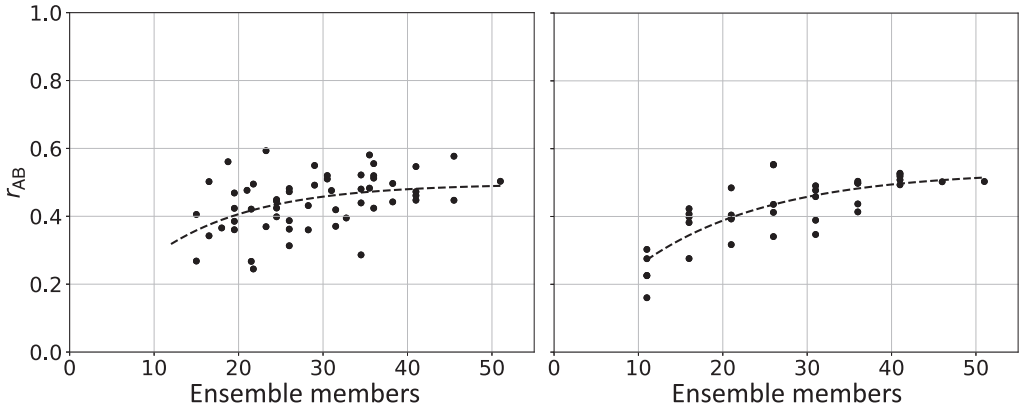
number of randomly selected perturbed members from each of the ECMWF and JMA ensembles. These ensembles were chosen as they had the largest number of members. Five pairs of reduced-member ensembles were generated for ensemble sizes, including the control forecast, of 11, 16, 21, 26, 31, 36, and 41 members. In addition, a single 46-member case was generated and the full 51-member case was also included in the comparison. The resulting spatial correlation between each of the 37 pairs is plotted in Figure 9b and clearly demonstrates a relationship between the r_{AB} values and the ensemble size. A best-fit relationship (dashed line) of the form

$$r_{\mathrm{AB}} = \alpha(1 - \mathrm{e}^{\beta N}), \tag{20}$$

where α and β are constants, has been fitted to the data in Figure 9. In both cases, the 50-member comparison approaches the asymptotic value of r_{AB}. These results imply that, as ensemble size increases, the spread-based diagnostics are less sensitive to the statistical effects of sampling and should, therefore, be more representative of the atmospheric uncertainty and its growth.

4.4 | PV diagnostics

In this section, diagnostics calculated from PV_{250} using data from the ECMWF operational archive are compared with those obtained for Z_{250}. Although the PV values have not been calculated for a constant θ surface, PV_{250} is expected to provide a better approximation than Z_{250} to a tracer-like variable, for which $\mathrm{NonAdv}_X = 0$. Therefore, the comparison between the PV_{250} and Z_{250} provides an indication of whether the $NA_{Z_{250}}$ values are likely



(a) Full ensemble size for all models (b) Reduced ensemble size comparison

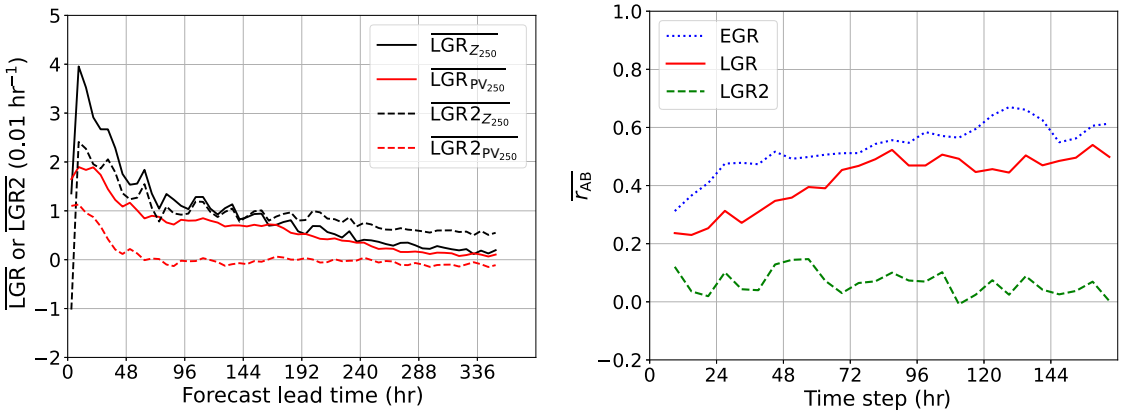
FIGURE 9 The effect of ensemble size on the spatial correlation r_{AB}, which compares the spatial distribution of diagnostic values for each ensemble pair, for the $LGR_{Z_{250}}$ diagnostic. Data are for northern midlatitudes and lead times between 48 hr and 72 hr from the forecast initiated at 0000 UTC May 1, 2022. (a) r_{AB} values for the full ensembles of all model pairs. (b) r_{AB} values for the reduced member ensembles obtained using equally sized randomly selected subsets of ensemble members from the European Centre for Medium-Range Weather Forecasts and Japan Meteorological Agency ensemble prediction systems. All data points are plotted against the average ensemble size of the respective ensemble prediction system pair.

to be significant in the comparisons between ensemble prediction systems.

Figure 10a compares the evolution of the $\overline{\text{LGR}}$ (solid lines) and the $\overline{\text{LGR2}}$ (dashed lines) values for the northern midlatitudes from the 0000 UTC May 1, 2022, forecast for diagnostics calculated from these two variables. The area-weighted average values for LGR and LGR2 are larger for Z_{250} than for PV_{250} throughout the forecast, although the difference for LGR becomes relatively small after the first 48-hr period. $\overline{LGR2_{PV_{250}}}$ is approximately zero after the first 48 hr, whereas $\overline{LGR2_{Z_{250}}}$ remains positive even when lead times exceed 300 hr. For each variable, short-term variations in the evolution of the $\overline{\text{LGR}}$ values are also present in the evolution of the $\overline{LGR2}$ values. Figure 10b compares the spatial correlation r between $EGR_{Z_{250}}$ and $EGR_{PV_{250}}$, between $LGR_{Z_{250}}$ and $LGR_{PV_{250}}$, and between $LGR2_{Z_{250}}$ and $LGR2_{PV_{250}}$. The r value for the EGR comparison increases from 0.3 at the start for forecast to a maximum of almost 0.7 after approximately 130 hr. There is a similar trend for the $LGR_{Z_{250}}$ diagnostics, although the values are slightly lower. In contrast to the good spatial correlations for the EGR and LGR diagnostics, there is very poor spatial correlation between the LGR2 diagnostics throughout the entire forecast. Similar characteristics to those shown in Figure 10 are present when comparing diagnostics for Z_{250} and PV_{250} throughout the midlatitude and polar regions and for all 12 forecasts investigated.

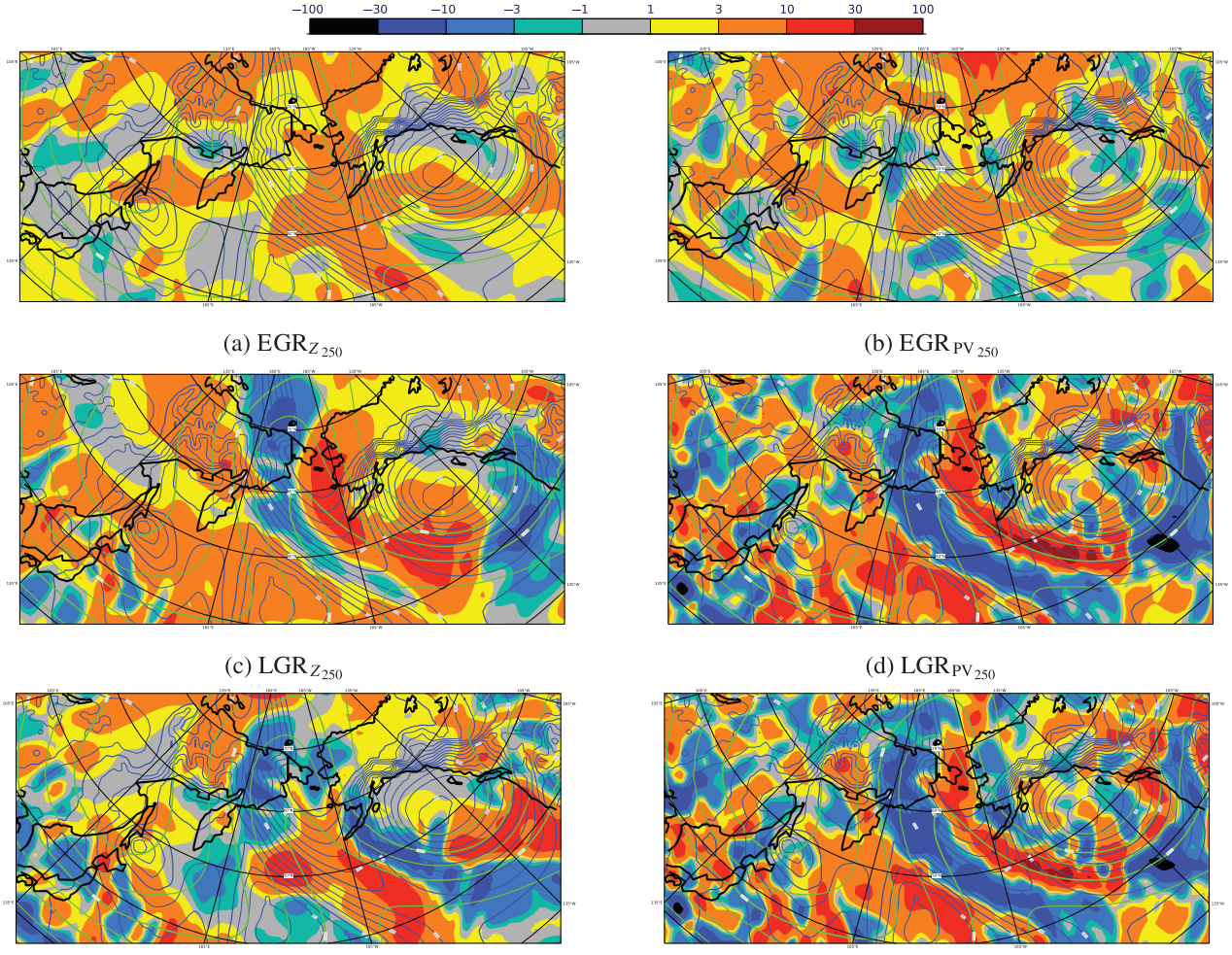
The spatial distributions of the EGR, LGR, and LGR2 diagnostic values for Z_{250} and PV_{250} at a lead time of 51 hr, and for the same region and forecast date as Figure 4, are shown in Figure 11. The regions of high and low EGR and LGR values calculated using the two variables are in similar locations. In addition, there is little difference between the values of the LGR and LGR2 diagnostics when calculated using PV_{250}. However, there are significant differences between the locations of the high and low LGR and LGR2 values for Z_{250}. This qualitative comparison between the LGR and LGR2 values is supported by spatial correlation analysis (not shown) with the $LGR_{PV_{250}}$ and $LGR2_{PV_{250}}$ strongly correlated ($r > 0.9$ for the first 90 hr of the forecast) whereas $LGR_{Z_{250}}$ and $LGR2_{Z_{250}}$ are poorly correlated ($r < 0.15$ after the first 48 hr of the forecast). Although, for PV, the $D_{PV_{250}}$ term is diffusive throughout the forecast, leading to the smaller area-weighted average for $LGR2_{PV_{250}}$ than for $LGR_{PV_{250}}$ seen in Figures 10a, 11d,f shows that the difference in these average values is generally small relative to the magnitude of the diagnostic values at any particular location. In addition, when added to the $LGR_{PV_{250}}$ values, the $D_{PV_{250}}$ term does not affect the location of the high and low values of the Lagrangian growth-rate diagnostics. In contrast, for Z, the $D_{Z_{250}}$ term does have a significant impact on the location of the high and low values of the Lagrangian growth-rate diagnostics, and is positive on average for lead times in excess of 144 hr.

The relatively good correlation between the EGR diagnostic for the two variables is consistent with the Z_{250} field, representing a smoothed version of the PV_{250} field. As advection is calculated using the same ensemble mean velocity for both variables, the LGR diagnostic is similarly consistent, and the spatial distribution of $LGR_{Z_{250}}$ can, therefore, be used as an approximation to $LGR_{PV_{250}}$. This is not the case for the LGR2 diagnostic, where the $D_{Z_{250}}$



(a) $\overline{\mathrm{LGR}}$ and $\overline{\mathrm{LGR2}}$ values (b) r values

FIGURE 10 Comparing the evolution of uncertainty growth-rate diagnostics calculated with Z_{250} and PV_{250} in the northern midlatitude region using the ECMWF forecast from 0000 UTC May 1, 2022: (a) area-weighted average values for LGR and LGR2; (b) spatial correlation coefficient between growth-rate diagnostics calculated for Z_{250} and PV_{250}. [Colour figure can be viewed at wileyonlinelibrary.com]



(a) $\mathrm{EGR}_{Z_{250}}$ (b) $\mathrm{EGR}_{\mathrm{PV}_{250}}$

(c) $\mathrm{LGR}_{Z_{250}}$ (d) $\mathrm{LGR}_{\mathrm{PV}_{250}}$

(e) $\mathrm{LGR2}_{Z_{250}}$ (f) $\mathrm{LGR2}_{\mathrm{PV}_{250}}$

FIGURE 11 Smoothed EGR, LGR, and LGR2 values (10^{-2} hr^{-1}) for Z_{250} and PV_{250}, in a region covering part of the North Pacific and Arctic, 51 hr into the European Centre for Medium-Range Weather Forecasts forecast initiated at 0000 UTC January 1, 2022. Also shown are the mean-sea-level pressure from the respective control forecasts (blue contours) and the ensemble mean value of Z_{250} (green contours). [Colour figure can be viewed at wileyonlinelibrary.com]

term has a large impact on the distribution for Z. Combining these results means that there are substantial differences in the sum of non-conservative and non-advective terms ($NC_X + NA_X$) for these two variables, implying that $LGR2_{Z_{250}}$ has a large non-advective component. Based on this result, it is recommended that the use of the LGR2 diagnostic is restricted to tracer-like variables.

5 | SUMMARY AND CONCLUSIONS

A comparison of forecast uncertainty in the operational forecasts of 12 ensemble prediction systems, with data retrieved from the TIGGE archive, has been conducted to understand whether ensemble-spread-based diagnostics provide a useful measure for understanding the processes that are associated with the growth and transport of atmospheric uncertainty. Diagnostics investigated were the ensemble standard deviation, as a direct measure of forecast spread, and Eulerian (or local) and Lagrangian growth tendencies and growth rates, where the growth rates are normalised forms of the growth tendencies. In addition to a Lagrangian growth-rate diagnostic based on the mean ensemble velocity (LGR), an alternative version (LGR2) has been derived that includes an additional transport term. This alternative version is a measure of the uncertainty growth rate associated with non-conservative diabatic and frictional processes. It has been argued that, when growth-rate diagnostics are relatively independent of the ensemble prediction system, the effect of model errors is likely to be small and the development of these diagnostics can be attributed to the growth of atmospheric uncertainty. Under these circumstances, the diagnostics can then be used to understand the flow-dependent evolution of atmospheric uncertainty.

Diagnostics for 250 hPa geopotential height were calculated for the 12 forecasts that were initiated on the first day of each month in 2022. Diagnostics obtained from the ensemble prediction systems were compared qualitatively using maps of the diagnostic spatial distributions, and quantitatively using area-weighted average values and the spatial correlation between the diagnostic maps. As 250 hPa geopotential height is not a tracer-like variable, the calculated diagnostics may include a significant non-advective component. Therefore, diagnostics for 250 hPa PV, a variable expected to be almost tracer-like, were also calculated for the ECMWF ensemble prediction system. Comparisons between the diagnostics using 250 hPa PV and geopotential height diagnostics from the ECMWF ensemble prediction system were conducted to understand the importance of the non-advective term.

The main conclusions of the study are as follows:

1. In the extratropical regions (north of 30° N and south of 30° S) the values of each of the growth-rate diagnostics were consistent across all 12 ensemble prediction systems for a range of forecast lead times starting at 48 hr and ending at between 96 hr and 192 hr. In addition, for this range of lead times (the second period), the spatial distribution of the growth diagnostics is shown to be similar for all ensemble prediction systems. These results imply that, during this second period, the flow-dependent nature of the uncertainty growth is a property of the atmosphere and not specific to one ensemble prediction system.
2. Growth rates in the midlatitude part of the extratropical regions are approximately constant for this second period, implying that growth is exponential.
3. For lead times of less than 48 hr (the first period) there is less consistency in the diagnostic values calculated across the ensemble prediction systems. The results at short lead times are attributed to differences in the initial perturbation structure and spread across the ensemble prediction systems, the effect of which reduces over time as the perturbations align with the main features of the synoptic-scale flow. Less consistency is also found at longer lead times (the third period), and this result is attributed to differences in the NWP model design that impact the location and strength of the main synoptic features, combined with the effects of saturation, which also appear to be model specific.
4. In tropical regions, ensemble spread grows linearly and growth-rate diagnostics are more inconsistent across the ensemble prediction systems than for the extratropics at all lead times. These results were found when investigating uncertainty in both geopotential-height- and velocity-based variables, including kinetic energy. Short time-scale fluctuations in the area-weighted average growth rates in these regions follow a diurnal cycle, suggesting that the diagnostic values are linked to convective processes, with the highest values when convection is at its maximum. The ensemble prediction system dependence implies that growth-rate diagnostics may be a less useful measure for understanding the sources and transport of atmospheric uncertainty growth in the tropical regions than in the extratropical regions.
5. Although diagnostic values for all ensemble prediction systems are similar in the extratropics in the second period, a large ensemble size is required to avoid sampling affecting the estimate of uncertainty growth rate and its flow dependence.
6. It has been shown in its derivation that the LGR2 diagnostic represents the uncertainty growth rate associated with the non-conservative source term for tracer-like variables and should, therefore, provide a useful tool for

assessing the sources of uncertainty in future studies. The differences between the LGR2 results for geopotential height and PV are attributed to a significant non-advective component in the geopotential-height diagnostics and, therefore, it is recommended that the use of the LGR2 diagnostic is restricted to variables that have tracer-like properties.

Overall, this work demonstrates that ensemble-spread-based diagnostics, calculated from the 12 ensemble prediction systems, are consistent in the extratropical regions for a range of forecast lead times starting at 48 hr and ending at between 96 hr and 192 hr. Calculation of the LGR2 diagnostic from a variable that has tracer-like properties should enable the effects of non-conservative sources and sinks on uncertainty growth to be understood. In addition, the geopotential height results indicate that an ensemble size of approximately 50 members is sufficient to obtain consistent and, therefore, representative growth-rate diagnostics between ensemble prediction systems. The consistency of these results provides confidence in the use of these diagnostics for identifying sources of atmospheric uncertainty and assessing the synoptic systems and atmospheric processes that have the largest impact on medium-range predictive skill.

ACKNOWLEDGEMENTS

We thank the ECMWF for providing access to the ensemble prediction system data within the TIGGE and ECMWF operational archives, along with the team managing these archives and the centres contributing to TIGGE. TIGGE is an initiative of the World Weather Research Programme. Doug Wood would also like to thank the members of staff at the University of Reading and ECMWF that have provided encouragement, guidance, and challenge during the completion of this work, and the University of Reading for the Regional PhD Bursary that supports his studies. We would also like to acknowledge the helpful comments from the two anonymous reviewers.

CONFLICT OF INTEREST STATEMENT

The contact author has declared that none of the authors have any competing interests.

DATA AVAILABILITY STATEMENT

The data that support the finding of this study are available from within the TIGGE archive and ECMWF operational archive.

The data that support the findings of this study are openly available in the TIGGE archive at https://confluence.ecmwf.int/display/TIGGE.

ORCID

Douglas Wood https://orcid.org/0009-0008-3381-3266

REFERENCES

Baumgart, M. & Riemer, M. (2019) Processes governing the amplification of ensemble spread in a medium-range forecast with large forecast uncertainty. *Quarterly Journal of the Royal Meteorological Society*, 145(724), 3252–3270.

Buizza, R., Houtekamer, P., Pellerin, G., Toth, Z., Zhu, Y. & Wei, M. (2005) A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, 133(5), 1076–1097.

Froude, L.S. (2010) TIGGE: Comparison of the prediction of northern hemisphere extratropical cyclones by different ensemble prediction systems. *Weather and Forecasting*, 25(3), 819–836.

Hoskins, B.J., McIntyre, M.E. & Robertson, A.W. (1985) On the use and significance of isentropic potential vorticity maps. *Quarterly Journal of the Royal Meteorological Society*, 111(470), 877–946.

Judt, F. (2020) Atmospheric predictability of the tropics, middle latitudes, and polar regions explored through global storm-resolving simulations. *Journal of the Atmospheric Sciences*, 77(1), 257–276.

Leutbecher, M. & Palmer, T.N. (2008) Ensemble forecasting. *Journal of Computational Physics*, 227(7), 3515–3539.

Lewis, J.M. (2005) Roots of ensemble forecasting. *Monthly Weather Review*, 133(7), 1865–1885.

Lorenz, E.N. (1963) Deterministic nonperiodic flow. *Journal of Atmospheric Sciences*, 20(2), 130–141.

Lorenz, E.N. (1969) The predictability of a flow which possesses many scales of motion. *Tellus*, 21(3), 289–307.

Palmer, T. (2019) The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145, 12–24.

Park, Y.-Y., Buizza, R. & Leutbecher, M. (2008) TIGGE: Preliminary results on comparing and combining ensembles. *Quarterly Journal of the Royal Meteorological Society*, 134(637), 2029–2050.

Rodwell, M.J., Richardson, D.S., Parsons, D.B. & Wernli, H. (2018) Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bulletin of the American Meteorological Society*, 99(5), 1015–1026.

Rodwell, M.J. & Wernli, H. (2023) Uncertainty growth and forecast reliability during extratropical cyclogenesis. *Weather and Climate Dynamics*, 4(3), 591–615.

Swinbank, R., Kyouda, M., Buchanan, P., Froude, L., Hamill, T.M., Hewson, T.D. et al. (2016) The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, 97(1), 49–67.

Vallis, G.K. (2017) *Atmospheric and oceanic fluid dynamics*. Cambridge, UK: Cambridge University Press, p. 916.

How to cite this article: Wood, D., Gray, S., Methven, J. & Rodwell, M. (2026) How consistently do ensemble prediction systems represent the growth of atmospheric uncertainty?. *Quarterly Journal of the Royal Meteorological Society*, e70168. Available from: https://doi.org/10.1002/qj.70168