Examining model error in potential temperature and potential vorticity via weather forecasts at different lead times

ORIGINAL ARTICLE

Examining model error in potential temperature and potential vorticity via weather forecasts at different lead times

Oscar Martínez-Alvarado[1] | **Claudio Sánchez**[2]

[1]National Centre for Atmospheric Sciences, Reading, United Kingdom

[2]Met Office, Exeter, United Kingdom

Correspondence
Oscar Martínez-Alvarado, Department of Meteorology, University of Reading, Early Gate, Reading, Berkshire, RG6 6BB, United Kingdom
Email: o.martinezalvarado@reading.ac.uk

Funding information
OM-A's contribution was funded by the United Kingdom's Natural Environment Research Council through the National Centre for Atmospheric Sciences

The examination of model error is fundamental to improve weather forecasts at any time scale. Here, model errors for two forecast lead times (12, 24 h) at the grid-point level are analysed using (i) the total Eulerian changes in variables, such as potential temperature and potential vorticity (PV), both conserved under adiabatic, frictionless conditions; and (ii) Lagrangian diabatic tracers. The latter refines the Eulerian analysis by decomposing the total Eulerian changes into materially-conserved and diabatically-generated components. For both analyses the behaviour of a theoretical unbiased model, for which the only assumption is that forecast error is zero when averaged over a large number of cases, is used as a reference. Deviations from this theoretical behaviour are used to highlight conditions leading to large errors. The analyses are performed on a set of forecasts produced with the United Kingdom's Met Office Unified Model for a 25-day period during the NAWDEX (North Atlantic Waveguide and Downstream Impact Experiment) field campaign (16 September–22 October 2016). The Eulerian approach indicates that changes in potential temperature and PV are underestimated with respect to the theoretical behaviour of an unbiased model. The grid points with the largest changes in 12-h forecasts have the largest underestimation in the 24-h forecast, highlighting the im-

portance of the underestimation for the most dynamically and thermodynamically active grid points. The Lagrangian-tracer investigation reveals very large deviations from the theoretical behaviour of an unbiased model regardless of the level of Eulerian change, in particular for PV, and an unrealistic similarity in magnitude between parametrised diabatic changes of PV in the 24-h and 12-h forecasts. This is at odds with what would be otherwise required to obtain unbiased behaviour. Addressing the deviations from the behaviour of a theoretical unbiased model found in this work could be a step forward towards an operational unbiased model.

KEYWORDS
Model error, Eulerian flow description, Lagrangian flow description, diabatic processes, diabatic tracers, potential vorticity, potential temperature

1 | INTRODUCTION

Numerical Weather Prediction is an initial value problem where the numerical representation of our current understanding of the physical laws governing atmospheric processes is integrated to a given validation time. A *perfect forecast* can be conceived as one which predicts with 100% accuracy the future state of the atmosphere, i.e. for a perfect forecast, forecast error $\epsilon \equiv 0$. Due to the chaotic nature of atmospheric dynamics, which for example makes the atmosphere's evolution sensitive to intial contitions, routine perfect forecasts could be obtained if and only if both the numerical model and the initial conditions were perfect, i.e. if and only if the relevant laws of physics were completely known and their numerical representation was 100% accurate, and the initial conditions fed into the model were completely free of error. None of these conditions are or will ever be met in reality and therefore routine perfect weather forecasts are impossible to obtain.

Given the impossibility to obtain routine perfect weather forecasts, we ask whether it is at all possible to achieve *unbiased forecasts*, i.e. forecasts free of systematic error. The definition of unbiased forecasts can only be done in statistical terms. Thus, an unbiased forecast model can be defined as one for which

$$\langle \epsilon \rangle = 0, \tag{1}$$

where $\langle \cdot \rangle$ is the mean over a large number of forecast-analysis pairs. Evidently a perfect forecast model is also an unbiased model, but an unbiased model is not necessarily a perfect model. Besides the practical benefits of having unbiased forecasts, there are theoretical consideration for which having such a tool would also be desirable. For example, estimations of the intrisic limit of predictability of the atmosphere can only be carried out under the assumption of a perfect model (e.g. Selz, 2019). However, the atmosphere and a numerical model of the atmsphere are two different dynamical systems, and therefore model-based estimations of intrinsic predictability might not be valid for the actual

atmosphere. An unbiased model would ensure that forecast errors only arise from an accurate representation of the atmosphere's intrinsic variability and not from the tendency of the model to move towards its own climatology.

The constant improvement of operational forecast models has allowed these models to become virtually unbiased at forecasting certain aspects of the atmopsheric system. For example, the systematic underestimation of Rossby-wave ridge area in forecasts produced by the Met Office Unified Model (MetUM) (Gray et al., 2014) has virtually disappeared for lead times of up to seven days over the North Alantic and Europe with the upgrade of the model's dynamical core (Martínez-Alvarado et al., 2018). However, there are other aspects for which biases remain. For example, the systematic underestimation of tropopause potential vorticity (PV) isentropic gradient in forecasts produced by the MetUM (Gray et al., 2014) remains despite the dynamical core's upgrade (Martínez-Alvarado et al., 2018), leading to erroneous Rossby-wave propagation in the forecasts (Harvey et al., 2016). Biases in forecasts of upper-level Rossby waves are not exclusive of the MetUM. Similar biases in operational forecasts produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) and the National Centers for Environmental Prediction were identified by Gray et al. (2014), and in the Korean Meteorological Administration (using a different configuration of the MetUM) by Martínez-Alvarado et al. (2018). Related biases can also be identified using different diagnotics such as the object-based forecast verification approach by Giannakaki and Martius (2016), which showed that the ECMWF Integrated Forecast System underestimated the area and strength of Rossby waveguides. Associated with upper-level forecast errors, there are long-standing systematic errors in forecasts of atmospheric blocking, whose frequency tends to be underestimated in forecasts at medium-range lead times (7–14 days) (Matsueda, 2009; Martínez-Alvarado et al., 2018). The biases related to Rossby waves and blocking at short lead times remain and evolve as lead time increases towards the sub-seasonal range. Even though these biases improve as model resolution increases, they are also dependent on the model representation of physical processes (Quinting and Vitart, 2019).

If we knew the sources and development mechanisms of forecast errors, we could devise means of disabling the sources and inhibiting the development of forecast error. Therefore, understanding forecast errors in numerical weather prediction models is critical for the improvement of models' accuracy. The dynamics of forecast error in terms of PV can be described through the formulation of a forecast error tendency equation (Davies and Didone, 2013). Developing this apprach further, Baumgart et al. (2019) have shown that tropopause forecast error growth follows the three-stage error growth model of Zhang et al. (2007) (see also Selz and Craig (2015)). At the third and final stage the tropopause forecast error growth is chiefly determined by near-tropopause dynamical processes rather than errors in tropospheric baroclinc wave development (Baumgart et al., 2018, 2019).

From a dynamical systems perspective, forecast errors arise as a consequence of imperfections in the models' initial state (*initial condition error*), the nonlinearity of the atmospheric dynamics (*inherent predictabiltiy*) and the imperfect numerical representation of atmospheric processes in the model (*model error*). The present work is concerned with the assessment of the latter. Model error arises from errors in the model tendencies computed by the model's components, namely the dynamical core, i.e. the numerical discretisation and solution of the equations of motion, and the parametrisation of physical processes, which represent processes which are not explicitly resolved by the dynamical core alone. The effects of these processes can be described by the changes they produce on variables, such as potential temperature (θ) or potential vorticity (Q), which would be conserved under adiabatic conditions. The aim of this study is to indirectly assess model error, by contrasting the changes sustained by θ and Q in an operational forecast model against those expected in a theoretical unbiased model, i.e. a model whose forecasts satisfy (1). Rather than focussing on tendencies over a single model time step (7.5 min, representing the discrete version of a continuous time differential), we have chosen to study the changes in θ and Q over a finite time interval (of the order of 12 hours).

The methodology consists of the comparison between short (T+12) and long (T+24) forecasts for the changes in θ and Q under two descriptions. The long forecasts lead time is chosen to one day to avoid large drifts in the atmospheric

flow that may affect the comparison of tendencies with the short forecast. The first description is Eulerian, in which the variables under investigation are the total changes in θ and Q at a given grid point with respect to the variables' values at the start of the forecast at that same grid point. The second description is based on Lagrangian diabatic tracers which track changes in potential temperature (Martínez-Alvarado and Plant, 2014; Martínez-Alvarado et al., 2014; Martínez-Alvarado et al., 2016a,b) and PV (Stoelinga, 1996; Gray, 2006; Chagnon and Gray, 2009; Chagnon et al., 2013; Chagnon and Gray, 2015; Martínez-Alvarado et al., 2016a,b; Saffin et al., 2016, 2017) along trajectories following the resolved winds. We call this description the Lagrangian-tracer description. Under this description, the changes in θ and Q are computed with respect to the variables' values characterising the air parcel, currently located at the grid point of interest, at the start of the forecast (in general at a different location). Thus, the Lagrangian-tracer description allows the decomposition of the Eulerian changes in θ and Q into diabatically-generated and materially-conserved components. In both cases, the results obtained with the operational forecast model are compared against the behaviour expected from a theoretical unbiased model.

The methodology is applied to hindcasts produced for the North Atlantic Waveguide and Downstream Impact Experiment (NAWDEX, Schäfler et al., 2018). NAWDEX was a large international field campaign to investigate the importance of diabatic processes for the development, evolution and predictability of upper-level Rossby waves over the North Atlantic and for their impacts downstream. NAWDEX involved the collaboration of several institutions in Europe and North America and the coordinated use of four research aircraft during the observation period that took place between 17 September and 22 October 2016. Several weather systems were observed during this period, including warm conveyor belts (WCBs), atmospheric rivers, extratropical transition of tropical cyclones, tropopause polar vortices and long-lived atmospheric blocking.

The rest of the article is organised as follows: The methodology and data are described in Section 2; the results, presented separately for each description, are shown in Section 3, in which the relationship between descriptions is discussed. The conclusions of the study are given in Section 4.

2 | METHODOLOGY AND DATA

2.1 | Methodology

A method that has proven useful in the identification of systematic *forecast* error is the comparison of forecasts at different initialisation times to highlight the effect of modelled processes on the evolution of the flow. In this method, a value of the target variable at analysis time (T+0) for a given forecast is determined. This value can then be used as a reference to compare those obtained at other forecast lead times. Given the atmosphere's inherent unpredictability, the values at other forecast lead times in a single-member forecast (either a so-called deterministic forecast or a single member of an ensemble forecast) are not expected to match those at analysis time. However, if the model was unbiased, the expected value over a large number of cases would match analysis-time value. Deviations from this behaviour reveal systematic errors and the drift of the forecast model towards its own climatology (e.g. Martínez-Alvarado, 2014). This method has been used in the identification (Gray et al., 2014) and further study (Martínez-Alvarado et al., 2018) of systematic errors in forecasts of Rossby-wave ridge areas and isentropic gradient of PV at the tropopause.

The method is relatively simple to implement if the variable of interest is an instantaneous variable, whose values can be determined unambiguously for a given validation time. However, if the variable of interest depends not only on validation time, but also on forecast lead time, the computation of appropriate forecast values corresponding to a given forecast lead time becomes more challenging. This is the case when the variable of interest represents the cumulative

change undertaken by an atmospheric variable, as this change will depend not only on the time of measurement, but also on the time when the accumulation started. To address this issue, the variable changes are investigated at the grid-point level, using two alternative descriptions of the flow. In the first one, we focus on the total changes that the variables undergo, following a purely Eulerian description of the flow. In the second one, Lagrangian diabatic tracers are used to separate the total Eulerian changes into changes due to advection only and changes due to the combined effect of advection and the parametrisation of sub-grid-scale diabatic processes.

The Eulerian approach is related to the *initial tendencies* (Klinker and Sardeshmukh, 1992; Rodwell and Palmer, 2007; Klocke and Rodwell, 2014) and *analysis tendencies* (Mapes and Bacmeister, 2012) methods to evaluate numerical models in that those methods also look at the total variable changes at a grid-point level. In the initial tendencies method, average forecast error is equated to average initial tendencies; in the analysis tendencies method, analysis tendencies are equated to the negative of model physics tendency error. By identifying similar patterns between forecast errors or analysis tendencies and the parametrised physical tendencies directly output from the models, model error can be assessed. In this work, model error is detected by identifying variable changes in short (12-h) forecasts and the corresponding changes in long (24-h) forecasts. These changes are then compared against each other and against those expected from the behaviour of a theoretical unbiased forecast model. Deviations between the theoretical model and the operational model highlight conditions leading to large errors. The Eulerian method is complemented by the Lagrangian-tracer approach by adding details about the physical (advective, frictional, mixing or diabatic) and numerical processes that lead to the Eulerian changes in the numerical model. The Lagrangian-tracer method was used by Saffin et al. (2017) to investigate the effects of processes that affect tropopause sharpness, known to be increasingly underestimated as forecast lead time increases (Gray et al., 2014). In this work we use the method in a different way, by again comparing corresponding Lagrangian variable changes in short and long forecasts and contrasting these to the behaviour we would expect in a theoretical unbiased forecast model.

2.1.1 | Eulerian description

Let $\varphi_k^s = \varphi_k^s(\boldsymbol{x})$ denote a generic variable φ (either θ or Q in this work) at forecast time step k from forecast base time s at a given grid point $\boldsymbol{x}$. To acknowledge the temporally discrete character of a numerical forecast, times are given in terms of arbitrary, but carefully chosen time steps as follows: To recover the actual times we define a reference time t_0 and assume that the forecast is initialised every T_b time units. Thus, the actual forecast base time is given by $t_b = t_0 + sT_b$. Assuming that the forecast is output every T_f time units, the validation time is given by $t = t_b + kT_f = t_0 + kT_f + sT_b$. If φ_n^m and φ_p^q are valid at the same time, then

$$p = n + (m - q)\frac{T_b}{T_f}.$$

For simplicity, we assume that $T_b = T_f = T$ in which case $p = n + m - q$ for two forecasts valid at the same time. As it will be explained in Section 2.2, in this work T is set to be 12 h.

We can write φ_k^s in terms of the values of the variable at analysis time φ_0^s by introducing an increment computed through a numerical forecast model $\Delta\varphi_k^s$ so that (see Fig. 1)

$$\varphi_k^s = \varphi_0^s + \Delta\varphi_k^s. \tag{2}$$

The term φ_0^s in (2) remains constant at each grid point during the forecast length. Note that, by definition, $\Delta\varphi_0^s \equiv 0$,

i.e. the forecast increment at the start of the forecast is zero . Forecast error ϵ_k^s can be defined through

$$\varphi_0^{s+k} = \varphi_k^s + \epsilon_k^s. \tag{3}$$

Note that $\epsilon_0^s \equiv 0$, i.e. the forecast error at the start of the forecast is zero.

We seek a relationship between $\Delta\varphi_{k+1}^s - \Delta\varphi_k^s$, i.e. the increment during one time step in a given forecast, and the first increment in a shorter forecast, $\Delta\varphi_1^{s+k}$, where the forecasts are $k > 0$ time steps apart. This objective can be achieved by writing the analysis valid at time $n + s + k$ in terms of two forecasts, namely φ_n^{s+k} and φ_{n+k}^s, as follows:

$$\varphi_0^{n+s+k} = \varphi_n^{s+k} + \epsilon_n^{s+k} = \varphi_{n+k}^s + \epsilon_{n+k}^s. \tag{4}$$

The forecasts can then be expanded using (2) to yield

$$\varphi_0^{n+s+k} = \varphi_0^{s+k} + \Delta\varphi_n^{s+k} + \epsilon_n^{s+k} = \varphi_0^s + \Delta\varphi_{n+k}^s + \epsilon_{n+k}^s. \tag{5}$$

Similarly, for the subsequent forecast time step $n + 1$

$$\varphi_0^{n+s+k+1} = \varphi_0^{s+k} + \Delta\varphi_{n+1}^{s+k} + \epsilon_{n+1}^{s+k} = \varphi_0^s + \Delta\varphi_{n+k+1}^s + \epsilon_{n+k+1}^s. \tag{6}$$

Subtracting (6) from (5) and making $n = 0$ yields

$$(\Delta\varphi_{k+1}^s - \Delta\varphi_k^s) - \Delta\varphi_1^{s+k} = \epsilon_1^{s+k} - \epsilon_{k+1}^s + \epsilon_k^s. \tag{7}$$

Equation (7) gives a relationship between the change in φ between two consecutive steps k and $k+1$ in a long forecast (with forecast base time s) and the change during the first step in a shorter forecast (with forecast base time $s + k$). This relationship is illustrated in Fig. 1. Taking the mean of (7) over a large number of cases and assuming that the forecast error mean is zero (unbiased-model assumption), we have

$$\langle \Delta\varphi_{k+1}^s - \Delta\varphi_k^s \rangle - \langle \Delta\varphi_1^{s+k} \rangle = 0, \tag{8}$$

where $\langle \cdot \rangle$ denotes the mean over a large number of cases. Equation (8) shows that on average the changes in φ between two consecutive validation times ($k + s$ and $k + s + 1$) should be the same for a forecast that just started ($\langle \Delta\varphi_1^{s+k} \rangle$) and one that has been running for longer ($\langle \Delta\varphi_{k+1}^s - \Delta\varphi_k^s \rangle$) if the model was unbiased. In contrast, if the model was biased, then the right-hand side of (8) would not be zero, indicating a systematic mismatch between the changes in variable φ for the same interval between two forecasts of different length. Following Leutbecher and Palmer (2008) we note that (8) is satisfied by an unbiased model for suficiently large subsamples conditioned on the size of the changes in the short-forecast.

2.1.2 | Lagrangian-tracer description

Diabatic tracers are sets of tracers describing the changes in θ (e.g. Martínez-Alvarado and Plant, 2014; Martínez-Alvarado et al., 2014) and Q (e.g. Stoelinga, 1996; Gray, 2006; Chagnon and Gray, 2009; Chagnon et al., 2013; Chagnon and Gray, 2015; Saffin et al., 2016) due to parametrised diabatic processes in a Lagrangian sense. These tracers have

been implemented in the Met Office Unified Model (MetUM, Walters et al., 2017). Tracers for Q have been used to study the decay in the sharpness of the tropopause by Saffin et al. (2017), while tracers for θ and Q have been used for the study of the development of forecast error in a case study by Martínez-Alvarado et al. (2016b) and for the comparison of the evolution of extratropical cyclones (Martínez-Alvarado et al., 2016b). Diabatic tracers for a given variable can be classified in two types. The first type, materially-conserved tracers, are affected by advection only and function as Lagrangian labels for the parcels in a Q-θ space. The second type, diabatically-generated tracers, are affected by advection and by local modifications due to parametrised tendencies. Thus, the variable φ^s_k can be written in terms of diabatic tracers as

$$\varphi^s_k = \varphi^s_{0,k} + \delta\varphi^s_k, \tag{9}$$

where $\varphi^s_{0,k}$ is conserved following trajectories and thus serves as a Lagrangian label for the trajectory and the air parcel, and $\delta\varphi^s_k$ represents the cumulative effect of diabatic, frictional and other parametrised processes on the air parcel. Notice that even though we refer to these changes as diabatic in the rest of the paper, they do include these other effects. While diabatic tracers have been described previously in e.g. Martínez-Alvarado et al. (2016a) the notation here is different from that in previous work to accommodate for forecasts with different base times. Comparing their Eq. (1) with (9), their φ, φ_0 and φ_d intialised at base time $t_b = t_b(s)$ and evaluated at valid times $t = t(k, s)$ become our φ^s_k, $\varphi^s_{0,k}$ and $\delta\varphi^s_k$, respectively. Note that their Eq. (1) includes a residual term r_φ, which for the forecast lead times we are using here (up to 24 hours) remains small and can be neglected (Martínez-Alvarado et al., 2016a). Thus, computing $\varphi^s_{0,k}$ and $\delta\varphi^s_k$, requires the solution of the governing equations (Martínez-Alvarado et al., 2016a)

$$\frac{D\varphi_0}{Dt} = 0 \quad \text{and} \quad \frac{D\varphi_\mathrm{d}}{Dt} = S_\varphi, \tag{10}$$

with initial conditions $\varphi_0(t_b(s)) = \varphi^s_0$ and $\varphi_\mathrm{d}(t_b(s)) = 0$, where S_φ represents diabatic and frictional sources of φ. These equations are solved within the MetUM, using the same numerical machinery that is used by the model to solve the evolution equations of its prognostic variables (Davis et al., 1993; Wood et al., 2014; Walters et al., 2017).

By definition, $\delta\varphi^s_0 \equiv 0$, i.e. the diabatically-generated tracer at the start of the forecast is zero. Consequently, $\varphi^s_0 \equiv \varphi^s_{0,0}$, i.e. the materially-conserved tracer matches the variable at analysis time at the start of the forecast. The structure of (9) is similar to that of (2). However, there are fundamental differences. Unlike φ^s_0 in (2), which remains constant at each grid point during the forecast, $\varphi^s_{0,k}$ in (9) varies as new air masses are advected into a particular grid point as the forecast evolves. Unlike $\Delta\varphi^s_k$, which represents the accumulated changes in φ in a given grid point from its value at the start of the forecast, $\delta\varphi^s_k$ represents changes in φ within the air parcel, which having started at a different location $\boldsymbol{x_0}$ is currently at the grid point under study at position $\boldsymbol{x}$.

Equating (2) and (9) we find the relationship between $\Delta\varphi^s_k$ and $\delta\varphi^s_k$:

$$\Delta\varphi^s_k = \delta\varphi^s_k + \varphi^s_{0,k\to 0}, \tag{11}$$

where $\varphi^s_{0,k\to 0} = \varphi^s_{0,k} - \varphi^s_0$ represents the replacement of the value φ^s_0, at a given grid point at the start of the forecast, by the value $\varphi^s_{0,k}$, advected by the resolved winds to be at the given grid point at the forecast time step k. Indeed, if the atmosphere was frictionless and adiabatic then $\delta\varphi^s_k \equiv 0$, by definition. Therefore, $\Delta\varphi^s_k = \varphi^s_{0,k\to 0}$, by (11), and $\varphi^s_k = \varphi^s_{0,k}$, by (2). This short-hand notation can be generalised. Thus,

$$\varphi^s_{0,k\to l} = \varphi^s_{0,k} - \varphi^s_l, \; k > l, \tag{12}$$

represents the replacement of the value φ_l^s, at a given grid point at time step l, by the value $\varphi_{0,k}^s$, advected by the resolved winds to be at the given grid point at time step k.

Using (11), we can rewrite (7) as

$$(\varphi_{0,k+1\to k}^s - \varphi_{0,1\to 0}^{s+k}) + (\delta\varphi_{k+1}^s - \delta\varphi_1^{s+k}) = \epsilon_1^{s+k} - \epsilon_{k+1}^s + \epsilon_k^s. \tag{13}$$

where (9) and (12) have been used. This relationship is illustrated in Fig. 2. Taking the mean of (13) over a large number of cases and using the unbiased-model assumption, we find that for an unbiased model

$$\langle \varphi_{0,k+1\to k}^s - \varphi_{0,1\to 0}^{s+k} \rangle + \langle \delta\varphi_{k+1}^s - \delta\varphi_1^{s+k} \rangle = 0. \tag{14}$$

Both terms in the first bracket represent advective replacement between two consecutive time steps (from $s+k$ to $s+k+1$), but the first term refers to the forecast starting at s, while the second refers to the forecast starting at the later time $s+k$. Similarly, the terms in the second bracket represent cumulative changes due to parametrised diabatic processes with the accumulation taking place from time s in the first term and from time $s+k$ in the second. The terms in the first and second brackets in (14) will be called hereafter Advective Replacement Difference (ARD) and Diabatic Modification Difference (DMD), respectively. To refer to ARD or DMD for a particular variable, the relevant variable will appear in brackets immediately after, e.g. $\mathrm{ARD}(\theta) = \theta_{0,k+1\to k}^s - \theta_{0,1\to 0}^{s+k}$ and $\mathrm{DMD}(Q) = \delta Q_{k+1}^s - \delta Q_1^{s+k}$.

While $\mathrm{DMD}(\varphi)$ involves a difference between the modification of φ along trajectories, $\mathrm{ARD}(\varphi)$ involves a difference between the materially-conserved values of φ at the start of the trajectories. Thus, under frictionless and adiabatic conditions $\mathrm{DMD}(\varphi) = 0$ while $\mathrm{ARD}(\varphi) \neq 0$, correctly indicating that forecast errors would only stem from errors in the advection as represented in the forecast model (by the so-called dynamical core). To aid the physical interpretation of the more complex and more realistic case, in which friction and diabatic changes are allowed, i.e. $\delta\varphi_n^m \neq 0$, let us consider the case of a grid point in a theoretical *perfect* forecast model with *perfect* initial conditions, for which forecasts at any lead time coincide with the analyses at the corresponding validation times. This situation is illustrated in Fig. 3, which represents the same situation as that illustrated in Fig. 1 (and Fig. 2), but now the Eulerian and Lagrangian increments correspond to perfect-forecast conditions. Notice that in this case the Eulerian forecast (black line) passes through the analyses, while the Lagrangian parcels' evolutions (red and blue lines) follow the same path regardless of the temporal point at which they start. Let us consider the relationships between the states in the forecasts and analyses in Fig. 3. By writing down two alternative expressions for the difference in φ between the state at time $s+k$ and that at time $s+k+1$ we obtain

$$\delta\varphi_{k+1}^s + (\varphi_{0,k+1}^s - \varphi_0^{s+k}) = \delta\varphi_1^{s+k} + (\varphi_{0,1}^{s+k} - \varphi_0^{s+k}), \tag{15}$$

which after after reorganising terms becomes

$$(\varphi_{0,k+1}^s - \varphi_{0,1}^{s+k}) + (\delta\varphi_{k+1}^s - \delta\varphi_1^{s+k}) = 0, \tag{16}$$

which is the perfect-forecast version of (14). This equation can be interpreted as follows: If we stood at a given point in the atmosphere and followed a parcel that will occupy that point, the changes in the value of an atmospheric variable at that point will be due to two effects, namely the advection of the parcel and the changes it undergoes as it travels from its origin to the selected point. If we compare these two changes between weather forecasts of different lengths, their differences must remain in balance, i.e. if the difference in the changes due to advection (quantified by

the ARD) is positive, then the difference in the changes along the trajectory (quantified by the DMD) must be of the same magnitude, but negative, and vice versa. If in Fig. 3 the value at the start of the red trajectory at time s was closer to the value of the analysis at time $s+k+1$, the total diabatic modification $\delta\varphi^{s}_{k+1}$ would be smaller. If everything else remained the same, then both ARD(φ) and DMD(φ) would have increased in magnitude to compensate the effect of the changes at time s.

It is worth pointing out that while in the Eulerian description, (8) measures an error between a reference value (the change in θ or Q in the short forecasts) and a proxy (the change in θ or Q in the long forecast), in the Lagrangian-tracer description (14) does not measure any error. Instead, it represents a balance between terms: ARD and DMD are allowed not to be zero (on average) as long as one is the same size as the other, but with opposite sign (on average).

2.2 | Data

The data is taken from a dataset comprising forecasts produced using the MetUM General Atmospheric configuration version 6.1 (GA6.1, Walters et al., 2017) covering the field campaign period of NAWDEX (Schäfler et al., 2018) that took place between 17 September 2016 and 22 October 2016. The present study includes the 25-day period comprising the forecasts from 0000 UTC 20 September 2016 to 1200 UTC 14 October 2016 every 12 hours, which includes the three storyline sequences of trigger, interaction, development and high-impact weather in Europe described in Schäfler et al. (2018, see their Fig. 6). Accordingly, the reference time t_0 was set to 0000 UTC 20 September 2016 and $T = 12$ h, which is the minimum T for the available data. The long forecast is set to be the forecast starting $T = 12$ h after the start of the short forecast (i.e. $k = 1$ througout this work).

The forecasts include diabatic tracers of θ and Q, so that $\varphi^{m}_{0,n}$ and $\delta\varphi^{m}_{n}$ in (9) are part of the model's output, in addition to other more commonly used meteorological fields such as mean sea level pressure. The fields are output on a domain bounded by 80°W and 40°E in longitude and by 20°N and 80°N in latitude comprising 514×385 grid points on each vertical model level. The investigation here has been carried out using one set of ten vertical model levels ($31 \leq l_m \leq 40$, where m_l is model level index). These MetUM terrain-following model levels are nominally located between 6.8 km and 11.2 km, i.e. in the upper troposphere/lower stratosphere. Thus, the statistical robustness of the results are ensured by including around 2×10^6 grid points for each date included in the study, and around 9×10^7 grid points for the whole forecast series. Samples of this size ensure statistical significance in one-sample t tests by producing very small variances of the means therefore leading to a large test statistic (Wilks, 2011). This is indeed the case for all the results involving means presented here. However, there is a caveat in that statistical independence cannot be ensured given that the grid points are correlated in both space and time. In order to address this issue we have sub-sampled the original datasets to produce 100000 smaller samples with 1000 data values each, and used the bootstrap method to compute statistical significance. With this method we have confirmed that our results are statistically significant.

3 | RESULTS

3.1 | Eulerian description

The results arising from the Eulerian description are shown in Fig. 4, in which the two terms in angular brackets in (8) are plotted. Considering $T = 12$ h and $k = 1$, the terms in (8) are interpreted as follows: $\Delta\varphi^{s+k}_{1}$ is the 12-h Eulerian change in φ in a forecast initialised just 12 hours before validation time, and $\Delta\varphi^{s}_{k+1}$ and $\Delta\varphi^{s}_{k}$ are the 24-h and 12-h Eulerian changes in φ in a forecast initialised 24 hours before validation time, respectively. A common feature between

the Eulerian changes in the 12-h forecasts, $\Delta\theta_1^{s+k}$ and ΔQ_1^{s+k}, is that their marginal histograms are centred, and largely concentrated, around zero (hinted by Fig. 4, but not explicitly shown). Therefore, in order to reveal a deviation from the behaviour of an unbiased forecast model, represented by the identity line, the data has been binned in ten equally populated bands between the p-th and $(p+1)$-th deciles of $\Delta\varphi_1^{s+k}$, for $p = 0, 1, \ldots, 9$.

The only bin for which zero is included in the 95% confidence interval of $\langle \Delta\varphi^s_{k+1} - \Delta\varphi^s_k \rangle - \langle \Delta\varphi_1^{s+k} \rangle$ is when $\Delta\varphi_1^{s+k}$ is between its fifth and sixth deciles, for both $\varphi = \theta$ and $\varphi = Q$, according to the bootstrap method used to test statistical significance. However, for values of $\Delta\varphi_1^{s+k}$ between its second and eighth deciles, $\langle \Delta\varphi^s_{k+1} - \Delta\varphi^s_k \rangle$ falls so close to $\langle \Delta\varphi_1^{s+k} \rangle$ that it can be said that the forecast model behaves like an unbiased model, for both $\varphi = \theta$ (Fig. 4a) and $\varphi = Q$ (Fig. 4b). For values below the second decile or above the eighth decile, there is an underestimation of $\langle \Delta\varphi^s_{k+1} - \Delta\varphi^s_k \rangle$ as a function of $\langle \Delta\varphi_1^{s+k} \rangle$ with respect to the behaviour of an unbiased forecast model. Furthermore, the position of the first and third quartiles of $(\Delta\varphi^s_{k+1} - \Delta\varphi^s_k)$ conditioned on $\Delta\varphi_1^{s+k}$ indicates that the whole distributions are shifted towards the horizontal axis, showing the tendency of the forecast model to underestimate the change in θ and Q in 24-h forecasts with respect to 12-h forecasts. Since this occurs for large absolute values of $\Delta\varphi_1^{k+s}$, it can be argued that the deviation from the behaviour of an unbiased model occurs on the most dynamically and thermodynamically active grid points, thus having a larger influence on the future state of the atmosphere.

The results described so far, which include all the 12-h–24-h forecast pairs in the dataset, were found to be qualitatively similar to those obtained from a single 12-h–24-h forecast pair, regardless of the time within the period under study. This assertion is also valid for the results pertaining to the Lagrangian-tracer description (Section 3.2). This allows us to relate our findings to specific meteorological features in a case study, for which we investigate a single forecast pair ($s = 23$), corresponding to a 24-h forecast with base time 1200 UTC 1 October 2016 and a 12-h forecast with base time twelve hours later, i.e 0000 UTC 2 October 2016. The single forecast pair corresponds to the development phase of the 'Stalactite cyclone' (Maddison et al., 2019), which developed over the North Atlantic between 1 October and 4 October 2016 (Schäfler et al., 2018). The cyclone was observed during the NAWDEX Intensive Observation Period 6, which consisted of a coordinated flight of the Deutsches Zentrum für Luft- und Raumfahrt (DLR) Falcon 20 and the French Service des Avions Français Instrumentés pour la Recherche en Environnement (SAFIRE) Falcon 20 (Schäfler et al., 2018). The synoptic situation was characterised by a prominent ridge extending from Greenland to Scandinavia and northwards beyond Iceland (e.g. Fig. 5). The Stalactite cyclone itself was a very deep system, which reached its maximum intensity in terms of mean sea level pressure around 0600 UTC 3 October 2016, when it exhibited a central pressure of 956 hPa, according to the analysis from the Met Office. The results are shown for validation time 1200 UTC 2 October 2016 when the cyclone's central pressure was 962 hPa located around 55°N, 27°W, according to the Met Office analysis. The synoptic situation is illustrated in the maps shown in Fig. 5 by means of mean sea level pressure, at low levels, and by the dynamic tropopause, represented by the 320-K 2-PVU (1 PVU = 1 K mm^2 kg^{-1} s^{-1}) PV contour, at upper levels. Using these fields as a frame of reference, the location of those grid points that exhibit the largest magnitude of $\Delta\varphi_1^{s+k}$ can be tied to specific synoptic features.

The geographical distribution of the grid points for which $|\Delta\theta_1^{s+k}|$ is maximal is related to the location of the Rossby-wave troughs and ridges (Fig. 5(a,b)). For $\Delta\theta_1^{s+k} < 0$, the grid points are located around the upstream trough, mainly on the stratospheric side (Fig. 5a); for $\Delta\theta_1^{s+k} > 0$, the grid points are located around the eastern edge of the downstream ridge, mainly on the tropospheric side (Fig. 5b).

Like in the case of θ, in the case of Q the dynamic tropopause is the synoptic feature that provides a reference to understand the geographical distribution of grid points for which $|\Delta Q_1^{s+k}|$ is maximal (Fig. 5(c,d)). Regardless of whether ΔQ_1^{s+k} is positive or negative, the grid points tend to be aligned along the dynamic tropopause. For $\Delta Q_1^{s+k} < 0$, the grid points tend to be located mainly on the tropospheric side (Fig. 5c). These grid points correspond to locations where the ridge is growing, i.e. strongly reducing Q from stratospheric to tropospheric values at those grid points.

This highlights that in a ridge (tropospheric air), the model tends to produce negative potential voricity increments, whose magnitude is too small. However, there are also several points that appear away from the tropopause, mainly on the stratospheric side. For $\Delta Q_1^{s+k} > 0$, the grid points are located mainly on the stratospheric side (Fig. 5d). There are several of these grid points within troughs, as the stratospheric air mass, carrying high PV, replaces tropospheric air characterised by low PV at those locations. This shows that in a trough (stratospheric air) , the model tends to produce positive PV increments, whose magnitude is again too small.

The results from the Eulerian description show that the statistical mismatch between changes in 24-h and 12-h forecasts for both θ and Q is small around the most frequent values of the changes in the 12-h forecasts. This shows that model error is small most of the time for most grid points. However, it is not the most common values that matter the most for the evolution of the atmosphere. The extreme values are those that have a larger influence on atmospheric dynamics and it is there where the largest mismatch between the 24-h and 12-h forecasts occurs. PV offers the clearest illustration of this point (Fig. 4b). The most frequent value of ΔQ_1^{s+k} (in the 12-h forecasts) is zero. When this occurs, the model adequately produces very small values of $\Delta Q_{k+1}^{s} - \Delta Q_k^{s}$ (in the 12-h forecasts). However, when the changes in Q are expected to be large, the model tends to underestimate these changes. Larger changes in Q lead to larger effects on the atmosphere's state. At upper levels, those points for which changes in Q are large are concentrated around Rossby-wave troughs and ridges (Fig. 5(c,d)), which are important features for the subsequent development of the Rossby waves themselves (e.g. Davies and Didone, 2013; Baumgart et al., 2018), for the development of other synoptic scale feature, such as precipitation (Martínez-Alvarado et al., 2018), and for downstream effects on the surface (e.g. Piaget et al., 2015). Therefore, a systematic underestimation of these large changes (larger *model* errors) could be the source of large *forecast* errors, such as forecast busts (e.g. Rodwell et al., 2013; Grams et al., 2018). As described by Schäfler et al. (2018), the period between 29 September and 3 October 2016 was one of three periods of reduced forecast skill during the NAWDEX field campaign. A natural question to ask is "What is the origin of the Eulerian discrepancy between the 12-h and the 24-h forecasts?" The answer is related to the origin of forecast error via (7). For upper-level Rossby waves, forecast error is closely related to the way in which θ and Q are modified within extratropical cyclones' WCBs and how this process is represented in numerical models (Martínez-Alvarado et al., 2016b; Baumgart et al., 2019). We can then hypothesise that the Eulerian discrepancy arises at least in part from the representation of WCBs (Grams et al., 2018) and possibly other mesoscale systems, such as mesoscale convective systems (e.g. Rodwell et al., 2013) and tropical cyclones undergoing tropical transition (Grams and Archambault, 2016, e.g.), with the ability to produce sufficiently large latent heat release to modify the upper tropospheric environment.

3.2 | Lagrangian-tracer description

In Section 3.1 we have shown that the largest mismatch between changes in 24-h and 12-h forecasts occur when the 12-h-forecast changes are large. We use the Lagrangian-tracer description to shed new light onto these findings, by presenting the analysis for those grid points for which $\Delta\varphi_1^{s+k}$ is below its first decile, between its fourth and sixth deciles and above its ninth decile. These grid points correspond to those in the leftmost, the two central and the rightmost bins in Fig. 4.

The results of the Lagrangian-tracer analysis are shown in Fig. 6. Model assessment under the Lagrangian-tracer description involves a balance relationship between changes due to the materially-conserved tracers and changes due to parametrised diabatic processes. This relationship is given by (14), which states that, if the forecast model is unbiased, the Advective Replacement Difference ($\mathrm{ARD}(\varphi) = \varphi_{0,k+1\rightarrow k}^{s} - \varphi_{0,1\rightarrow 0}^{s+k}$) is on average equal in magnitude, but opposite in sign, to the corresponding Diabatic Modification Difference ($\mathrm{DMD}(\varphi) = \delta\varphi_{k+1}^{s} - \delta\varphi_1^{s+k}$). This is the

relationship that we shall test in this Section. The test is carried out by binning the data in ten equally populated bands between the p-th and $(p+1)$-th deciles of ARD(φ), for $p = 0, 1, \ldots, 9$ to reveal tendencies dependent on the magnitude of these differences.

The behaviour of $\langle \mathrm{DMD}(\theta) \rangle$ as a function of $\langle \mathrm{ARD}(\theta) \rangle$ depends on the $\Delta\theta_1^{s+k}$ bin (Fig. 6(a–c)). For the two central $\Delta\theta_1^{s+k}$ bins (Fig. 6b), the behaviour is close to the theoretical unbiased forecast model, which is consistent with the Eulerian findings. However, for the most extreme $\Delta\theta_1^{s+k}$ bins (Fig. 6(a, c)), the DMD(θ) values are underestimated by up to 2.5 K with respect to those required by the unbiased-model assumption for grid points above the first decile of ARD(θ). Furthermore, the whole distribution of DMD conditioned on ARD also exhibit a very strong underestimation, so that the theoretical mean behaviour of an unbiased model almost always falls outside the interval between the first and third quartiles of DMD(θ). The deviation is especially noticeable for the grid points for which ARD(θ) > 0.

The deviation from the theoretical behaviour of an unbiased model is even larger in the case of Q (Fig. 6(d–f)). There are only bins for which zero is included in the 95% confidence interval of $\langle \mathrm{DMD}(Q) \rangle + \langle \mathrm{ARD}(Q) \rangle$ is when ARD(Q) is between its third and fourth deciles in Fig. 6d, between its fourth and fifth deciles in Fig. 6e, and between its sixth and seventh deciles in Fig. 6f, according to the bootstrap method used to test statistical significance. $\langle \mathrm{DMD}(Q) \rangle$ as a function of $\langle \mathrm{ARD}(Q) \rangle$ describe lines with slopes between -0.11, for $\Delta\theta_1^{s+k}$ above its ninth decile (Fig. 6f), and -0.37, for $\Delta\theta_1^{s+k}$ between its fourth and sixth deciles (Fig. 6e). These slopes are much greater than the slope of -1 expected from an unbiased model, and lead to deviations with respect to this model of more than 3 PVU, for the most extreme ΔQ_1^{s+k} bins (Fig. 6f). Even though the slope of $\langle \mathrm{DMD}(Q) \rangle$ as a function of $\langle \mathrm{ARD}(Q) \rangle$ is small with respect to that of the unbiased model for the central ΔQ_1^{s+k} bins (Fig. 6e), the ARD(Q) values are located closer to zero than those in the extreme ΔQ_1^{s+k} bins (i.e. $|\mathrm{ARD}(Q)| < 1$ PVU). This limits the magnitude of the deviation with respect to the behaviour of an unbiased model to around 0.4 PVU (for ARD(Q) $= \pm 0.5$ PVU), which is consistent with the Eulerian findings. Moreover, the distribution of DMD(Q) conditioned on ARD(Q) exhibits a noticeable underestimation of the whole distribution with respect to the unbiased case, as indicated by the positions of the first and third quartiles of DMD(Q). These results show that the differences in diabatic modification between the 24-h and 12-h forecasts should have been much larger in order to match the magnitude of the differences in the changes due to advection, as required for the forecasts to be unbiased (See Fig. 3 and its discussion in Section 2.1.2).

3.3 | Combined effects of deviations in θ and Q and relationship to Eulerian description

Up to this point we have analysed the deviations in θ and Q separately. To show the combined effect of these deviations on both variables, we turn again to the Stalactite cyclone as a case study and we concentrate on those grid points that exhibit maximum deviation from the behaviour of a theoretical unbiased model. Thus, we restrict the analysis to those grid points for which the magnitudes of the Eulerian increments $\Delta\theta_1^{s+k}$ and ΔQ_1^{s+k} are greater than their respective sixth decile. Given the symmetry of the distribution of $\Delta\varphi_1^{s+k}$, for both $\varphi = \theta$ and $\varphi = Q$ (see Fig. 4), by using this threshold we are essentially selecting the same grid points as those shown in Figs. 5(a,b) and 5(c,d) for θ and Q, respectively. We also only include points for which ARD(Q) is greater than its eighth decile, as these exhibit the maximal Lagrangian-tracer deviation as illustrated in Fig. 6(d–f). Furthermore, we restrict the data to the troposphere only (i.e. grid points for which $Q < 2$ PVU at the end of the 12-h forecast). These grid points are mainly concentrated within the large-amplitude Rossby-wave ridge, upstream of the Stalactite cyclone. The number of grid points per column satisfying these conditions in the 10-level column (recall $31 \leq m_l \leq 40$) is close to 10 towards the ridge's eastern flank (Fig. 7). Taking the average over these grid points for $\varphi_{0,1\to 0}^{s+1}$ and $\delta\varphi_1^{s+1}$ (corresponding to the 12-h forecast) and using (9), we can compute the average Eulerian increments between the two consecutive time steps $s+k$ and $s+k+1$

TABLE 1 Average terms in the Stalactite cyclone case study (see text for specification of data points).

		$\langle \theta^{s+k}_{0,1\to 0} \rangle > 0$		$\langle \theta^{s+k}_{0,1\to 0} \rangle < 0$	
		θ (K)	Q (PVU)	θ (K)	Q (PVU)
12-h advective replacement	$\langle \varphi^{s+k}_{0,1\to 0} \rangle$	5.55	-2.43	-5.64	-4.64
12-h diabatic modification	$\langle \delta\varphi^{s+k}_{1} \rangle$	-0.90	-0.04	-0.26	-0.07
ARD(φ)	$\langle \varphi^{s}_{0,k+1\to k} - \varphi^{s+k}_{0,1\to 0} \rangle$	-4.48	1.23	-0.03	1.43
DMD(φ)	$\langle \delta\varphi^{s}_{k+1} - \delta\varphi^{s+k}_{1} \rangle$	4.20	-0.29	0.45	-0.31
Normalised Eulerian difference	$\frac{\langle \Delta\varphi^{s}_{k+1} - \Delta\varphi^{s}_{k} \rangle - \langle \Delta\varphi^{s+k}_{1} \rangle}{\langle \Delta\varphi^{s+k}_{1} \rangle}$	-0.06	-0.38	-0.07	-0.24

for the 12-h forecast as

$$\Delta\varphi^{s+k}_{1} = \varphi^{s+k}_{0,1\to 0} + \delta\varphi^{s+k}_{1}. \tag{17}$$

By additionally taking the average over those same grid points for ARD and DMD and considering the definitions of these two diagnostics, we can compute the corresponding Eulerian increments for the 24-h forecast as

$$\Delta\varphi^{s}_{k+1} - \Delta\varphi^{s}_{k} = \Delta\varphi^{s+k}_{1} + \mathrm{ARD}(\varphi) + \mathrm{DMD}(\varphi), \tag{18}$$

where (17) has been used. Equation (18), obtainable also by equating (7) and (13), provides the link between the Eulerian and the Lagrangian-tracer descriptions, by explicitly showing that the imbalance between ARD and DMD give rise to the error between the Eulerian changes in the 12-h and the 24-h forecasts.

For the grid points used in this part of the study, the advective replacement in Q in the 12-h forecasts is generally negative, i.e. the parcel at a given location at time $s + k + 1$ is characterised by a Q-value at the start of the 12-h forecast (at time $s + k$), which is generally lower than the Q-value of the air parcel at the same location at time $s + k$. The corresponding replacement in θ can be either positive or negative, i.e. the parcel at a given location at time $s+k+1$ is characterised by a θ-value at the start of the 12-h forecast (at time $s + k$), which can be either lower or higher than the θ-value of the air parcel at the same location at time $s + k$. We present results for these two alternatives in Table 1. The first two data columns in Table 1 correspond to a positive advective θ-replacement. In this case, the differences in the Eulerian increments, 24-h minus 12-h forecasts, normalised by the increments in the 12-h-forecast are -0.06 for θ and -0.38 for Q. The last two data columns in Table 1 correspond to a negative advective θ-replacement. In this case, the differences in the Eulerian increments, 24-h minus 12-h forecasts, normalised by the increments in the 12-h forecast are -0.07 for θ and -0.24 for Q.

These results show that for θ the balance between the advective replacement and the changes due to parametrised processes produce similar Eulerian θ-increments in the last 12-h periods in both forecasts. By contrast, the relative difference between Eulerian Q-increments is much larger. This is a direct effect from the mismatch between forecasts demonstrated using the Lagrangian-tracer description. A potential explanation for these results is that the diabatic changes in the long forecast are too small. An alternative explanation is that the wrong parcel is being advected to these grid points, leading to an artificially inflated ARD(Q). Deciding which explanation is the correct one is not an easy task as advection and diabatic changes do influence each other (Martínez-Alvarado et al., 2016a). It is generally accepted that errors related to the dynamical core are small (e.g. Mapes and Bacmeister, 2012). If this is the case, then

the former explanation is correct. However, recent work has highlighted mismatch issues arising from differential advection depending on which variable is being advected (Whitehead et al., 2015; Saffin et al., 2016), which does not allow the ruling out of the alternative explanation. This is further supported by the forecast-bust case study by Grams et al. (2018), in which a forecast WCB which was too strong led to enhanced modification of PV at the WCB outflow region, suggesting the importance of both diabatic processes and advection in the development of forecast error.

4 | CONCLUSION

The evolution of potential temperature and PV from operational forecasts initialised at different times was studied under two descriptions. The first one is the Eulerian description, in which the investigation was focused on the total changes in potential temperature and PV at the grid-point level. The second description is based on Lagrangian tracers, which allow for the decomposition of the changes in two parts, a materially conserved part, which served as a Lagrangian label, and a diabatically generated part due to the combined action of the parametrised diabatic processes and advection.

The conceptual model that arises is rather complex, but we try to simplify it by considering the perfect-forecast case (which would require a perfect model and perfect initial conditions). In this case, the Eulerian analysis tells us that if we stood at a given point in the atmosphere and simulated the change an atmospheric variable would undergo in a particular time interval (with a defined start and end), this change would be the same regardless of when we started the simulation. Still considering the perfect-forecast case, the Lagrangian-tracer analysis tells us that, if we followed the parcel that will occupy the point in the atmosphere at which we are standing, the changes in the value of an atmospheric variable will be due to two effects: the advection of the parcel and the changes it undergoes as it travels from its origin to the selected point. If we compare these two changes between weather forecasts of different length, their differences must remain in balance. Thus, if the changes due to advection are smaller in the short forecast than in the long forecast, then the changes along the trajectory must be larger in the short forecast than in the long forecast and vice versa. We have formalised these relationships by introducing the concepts of Advective Replacement Difference ($\mathrm{ARD}(\varphi) = \varphi^{s}_{0,k+1\to k} - \varphi^{s+k}_{0,1\to 0}$) and Diabatic Modification Difference ($\mathrm{DMD}(\varphi) = \delta\varphi^{s}_{k+1} - \delta\varphi^{s+k}_{1}$), defined through (14) and the discussion that follows that equation.

As we explicitly state in Section 1, obtaining a perfect forecast is not possible and therefore we assume a less restrictive unbiased-forecast scenario and ask how close the behaviour of a state-of-the-art forecast model is to that of the theoretical unbiased model. Thus, the statistical expressions that we present in this work disregard the unrealistic expectation of a perfect forecast and lessen the constraints by considering instead the unbiased-forecast case. These unbiased-model relationships were tested on a dataset of 12-h and 24-h forecasts initialised at 00Z and 12Z from a 25-day period during September-October 2016, corresponding to the NAWDEX field campaign (Schäfler et al., 2018).

Using the Eulerian description, it was found that the operational forecast model tends to produce changes in the 24-h forecast which underestimate the corresponding changes in the 12-h forecast. This effect was displayed by both potential temperature and PV, and in both cases, the largest underestimation took place on the most dynamically and thermodynamically active regions characterised by the largest changes in both variables. In this study the regions of large changes in θ and Q corresponded to the location of Rossby-wave troughs and ridges, which are known to be important for the downstream development of these waves themselves (e.g. Davies and Didone, 2013; Baumgart et al., 2018) and have been linked to the occurrence of forecast busts (Rodwell et al., 2013; Grams et al., 2018). Forecast error in these regions is closely related to the way in which θ and Q are modified within extratropical cyclones' WCBs

and how this process is represented in numerical models (Martínez-Alvarado et al., 2016b; Baumgart et al., 2019). We have hypothesised that the underestimation in the Eulerian changes of θ and Q arises at least in part from the model's representation of systems with the ability to modify the mid-latitude upper tropospheric environment via latent heat release, such as WCBs (Grams et al., 2018), mesoscale convective systems (e.g. Rodwell et al., 2013) and tropical cyclones undergoing tropical transition (Grams and Archambault, 2016, e.g.). However, further work is needed to confirm this hypothesis.

The underestimation of changes in θ and Q found with the Eulerian description was further investigated using the Lagrangian-tracer description: For potential temperature, it was shown that small Eulerian changes were characterised by a Lagrangian behaviour closer to that of a theoretical unbiased model; however, large Eulerian changes were accompanied by large Lagrangian deviations from the unbiased model's behaviour, manifest as the underestimation of DMD(θ) for a given ARD(θ). For PV, it was shown that a clear underestimation of DMD(Q) for a given ARD(Q) with respect to the behaviour of a theoretical unbiased model was present regardless of the level of Eulerian change. Thus, the better approximation to the unbiased model's behaviour for small-magnitude Eulerian changes was due to these changes being associated to small-magnitude ARD(Q) rather than to a better Lagrangian behaviour per se.

In this work we have studied forecasts at a maximum lead time of 24 hours. Forecast error at these lead times is generally very small, leading to forecast skill possibly around 98% (See e.g. Fig. 1 in Bauer et al., 2015). On the other hand, forecast busts, i.e. occasional episodes of noticeably low forecast skill (e.g. Rodwell et al., 2013), occur at lead times of the order of five days. Therefore, connecting our results to episodes of large forecast error is not straightforward. However, we can hypothesise that the cumulative effect of the deviations from the behaviour of an unbiased model contribute to the growth of forecast error and, under certain circumstances, to the occurrence of forecast busts.

While the whole 25-day dataset was used to ensure the results' statistical robustness, it was found that single 12-h–24-h forecast pairs exhibit the same qualitative behaviour as the whole dataset. This suggests that deviations from the unbiased model's behaviour in the changes of potential temperature and PV do not depend on the synoptic situation, recalling that the NAWDEX field campaign period was characterised by a diversity of synoptic situations including extratropical transition of tropical cyclones, strong WCBs, tropopause polar vortices and atmospheric blocking (for a more complete account see Schäfler et al., 2018). Assuming a direct relationship between these changes and model error, understood as error in the model tendencies, these results suggest that the statistics of *model* error are flow-independent to a large extent in contrast to *forecast* error, which is widely known to be flow-dependent (e.g. Ferranti et al., 2015). In this work, we studied total changes in potential temperature and PV. A natural next step would be to study the separate effects of individual parametrisations and their interactions.

Given the flow-independence of the statistical results, we were able to use single 12-h–24-h forecast pairs as case studies. Thus, the Stalactite cyclone, which developed during the first days of October 2016, was used to illustrate the relationship between changes in potential temperature and PV and particular meteorological features. It was shown that most grid points exhibiting large Eulerian deviations from the unbiased model's behaviour were also part of dynamically and thermodynamically active regions at upper levels, such as Rossby-wave troughs and ridges. This reinforces the idea of the dynamical importance of the deviations from the unbiased model's behaviour for the accuracy of the forecasts and provides motivation for further investigation. Restricting the analysis to the troposphere, the Lagrangian-tracer analysis revealed that, while the Eulerian θ-increments in the 24-h forecast were close to those in the 12-h forecast, the corresponding Q-increments in the 24-h forecast were underestimated with respect to those in the 12-h forecast. This mismatch could arise from several potential sources. Given the location of these grid points, mainly along the Rossby-wave ridge, we hypothesise that the underestimation in the Q-increments could be related to the reduction in PV gradient as identified by Gray et al. (2014) and further studied by Harvey et al. (2016) and Saffin

et al. (2017). However, more work is needed to firmly establish this link. From a Langrangian point of view, the mismatch could be a consequence of an underestimation in the depletion of PV in the 24-h forecast, which would point to errors in the parametrisation of diabatic processes. Alternatively, it could result from the advection of the wrong parcels into the affected regions, which could yield inaccurate estimations in the 24-h-minus-12-h forecast differences. Determining the correct explanation can prove a challenging task as advection and parametrised diabatic processes are intimately related both in reality and within the machinery of numerical forecast models (Martínez-Alvarado et al., 2016a). The role of the intricate relationship between diabatic processes and advection has been demonstrated in case studies of large forecast error (e.g. Grams et al., 2018). We argue here that this is not exclusive of these episodes, but pervasive throughout the performace of the model and an expression of errors in model formulation. Thus, despite the challenge that explaining the mismatch between advection and diabatic modification in models poses, unveiling the relative importance of these factors and their systematic occurrence could lead to important improvements in forecast skill. Another important remaining unsolved aspect is the development of the deviations as forecasts progress. In this work we have investigated 12- and 24-hour forecasts starting every 12 hours, but it would be worth reducing both lead time and time interval to investigate how the deviations evolve.

Acknowledgements

The authors thanks Drs S. L. Gray and K. D. Williams for their comments to earlier versions of the manuscript and four anonymous reviewers for very useful comments to improve this work. Data from the simulations is archived at the Met Office and available for research use through the Centre for Environmental Data Analysis JASMIN platform (http://www.jasmin.ac.uk/); please contact the authors for details.

references

Bauer, P., Thorpe, A. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, **525**, 47–55.

Baumgart, M., Ghinassi, P., Wirth, V., Selz, T., Craig, G. C. and Riemer, M. (2019) Quantitative view on the processes governing the upscale error growth up to the planetary scale using a stochastic convection scheme. *Mon. Weather Rev.*, **147**, 1713–1731.

Baumgart, M., Riemer, M., Wirth, V., Teubler, F. and Lang, S. T. K. (2018) Potential vorticity dynamics of forecast errors: A quantitative case study. *Mon. Weather Rev.*, **146**, 1405–1425.

Chagnon, J. M. and Gray, S. L. (2009) Horizontal potential vorticity dipoles on the convective storm scale. *Q. J. R. Meteorol. Soc.*, **135**, 1392–1408.

— (2015) A diabatically-generated potential vorticity structure near the extratropical tropopause in three simulated extratropical cyclones. *Mon. Weather Rev.*, **143**, 2337–2347.

Chagnon, J. M., Gray, S. L. and Methven, J. (2013) Diabatic processes modifying potential vorticity in a North Atlantic cyclone. *Q. J. R. Meteorol. Soc.*, **139**, 1270–1282.

Davies, H. C. and Didone, M. (2013) Diagnosis and dynamics of forecast error growth. *Mon. Weather Rev.*, **141**, 2483–2501.

Davis, C. A., Stoelinga, M. T. and Kuo, Y.-H. (1993) The integrated effect of condensation in numerical simulations of extratropical cyclogenesis. *Mon. Weather Rev.*, **121**, 2309–2330.

Ferranti, L., Corti, S. and Janousek, M. (2015) Flow-dependent verification of the ECMWF ensemble over the Euro-Atlantic sector. *Q. J. R. Meteorol. Soc.*, **141**, 916–924.

Giannakaki, P. and Martius, O. (2016) An Object-Based Forecast Verification Tool for Synoptic-Scale Rossby Waveguides. *Wea. Forecasting*, **31**, 937–946.

Grams, C. M. and Archambault, H. M. (2016) The key role of diabatic outflow in amplifying the midlatitude flow: A representative case study of weather systems surrounding western North Pacific extratropical transition. *Mon. Weather Rev.*, **144**, 3847–3869. URL: https://doi.org/10.1175/MWR-D-15-0419.1.

Grams, C. M., Magnusson, L. and Madonna, E. (2018) An atmospheric dynamics perspective on the amplification and propagation of forecast error in numerical weather prediction models: A case study. *Q. J. R. Meteorol. Soc.*, **144**, 2577–2591.

Gray, S. L. (2006) Mechanisms of midlatitude cross-tropopause transport using a potential vorticity budget approach. *J. Geophys. Res.*, **111**, 14 pp.

Gray, S. L., Dunning, C., Methven, J., Masato, G. and Chagnon, J. (2014) Systematic model forecast error in Rossby wave structure. *Geophys. Res. Lett.*, **41**.

Harvey, B. . J., Methven, J. and Ambaum, M. H. P. (2016) Rossby wave propagation on potential vort icity fronts with finite width. *J. Fluid Mech.*, **794**, 775–797.

Klinker, E. and Sardeshmukh, P. D. (1992) The diagnosis of mechanical dissipation in the atmosphere from large-scale balance requirements. *J. Atmos. Sci.*, **49**, 608–627.

Klocke, D. and Rodwell, M. (2014) A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. *Q. J. R. Meteorol. Soc.*, **140**, 517–524.

Leutbecher, M. and Palmer, T. N. (2008) Ensemble forecasting. *J. Comput. Phys.*, **227**, 3515–3539.

Maddison, J. W., Gray, S. L., Martínez-Alvarado, O. and Williams, K. D. (2019) Upstream Cyclone Influence on the Predictability of Block Onsets over the Euro-Atlantic Region. *Mon. Weather Rev.*, **147**, 1277–1296.

Mapes, B. E. and Bacmeister, J. T. (2012) Diagnosis of tropical biases and the MJO from patterns in the MERRA analysis tendency fields. *J. Clim.*, **25**, 6202–6214.

Martínez-Alvarado, O. (2014) Implications of model error for numerical climate prediction. *Nonlin. Processes Geophys. Discuss.*, **1**, 131–153.

Martínez-Alvarado, O., Gray, S. L. and Methven, J. (2016a) Diabatic processes and the evolution of two contrasting summer extratropical cyclones. *Mon. Weather Rev.*, **144**, 3251–3276.

Martínez-Alvarado, O., Joos, H., Chagnon, J., Boettcher, M., Gray, S. L., Plant, R. S., Methven, J. and Wernli, H. (2014) The dichotomous structure of the warm conveyor belt. *Q. J. R. Meteorol. Soc.*, **140**, 1809–1824.

Martínez-Alvarado, O., Madonna, E., Gray, S. L. and Joos, H. (2016b) A route to systematic error in forecasts of Rossby waves. *Q. J. R. Meteorol. Soc.*, **142**, 196–210.

Martínez-Alvarado, O. and Plant, R. S. (2014) Parameterised diabatic processes in numerical simulations of an extratropical cyclone. *Q. J. R. Meteorol. Soc.*, **140**, 1742–1755.

Martínez-Alvarado, O., Maddison, J. W., Gray, S. L. and Williams, K. D. (2018) Atmospheric blocking and upper-level Rossby-wave forecast skill dependence on model configuration. *Q. J. R. Meteorol. Soc.*, **144**, 2165–2181.

Matsueda, M. (2009) Blocking predictability in operational medium-range ensemble forecasts. *SOLA*, **5**, 113–116.

Piaget, N., Froidevaux, P., Giannakaki, P., Gierth, F., Martius, O., Riemer, M., Wolf, G. and Grams, C. M. (2015) Dynamics of a local Alpine flooding event in October 2011: moisture source and large-scale circulation. *Q. J. R. Meteorol. Soc.*, **141**, 1922–1937.

Quinting, J. and Vitart, F. (2019) Representation of synoptic-scale rossby wave packets and blocking in the S2S prediction project database. *Geophys. Res. Lett.*, **46**, 1070–1078.

Rodwell, M. J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., Diamantakis, M., Earnshaw, P., Garcia-Mendez, A., Isaksen, L. et al. (2013) Characteristics of occasional poor medium-range weather forecasts for Europe. *B. Am. Meteorol. Soc.*, **94**, 1393–1405.

Rodwell, M. J. and Palmer, T. N. (2007) Using numerical weather prediction to assess climate models. *Q.J.R. Meteorol. Soc.*, **133**, 129–146.

Saffin, L., Gray, S. L., Methven, J. and Williams, K. D. (2017) Processes maintaining tropopause sharpness in numerical models. *J. Geophys. Res.: Atmos.*, **122**, 9611–9627. 2017JD026879.

Saffin, L., Methven, J. and Gray, S. L. (2016) The non-conservation of potential vorticity by a dynamical core compared with the effects of parametrized physical processes. *Q. J. R. Meteorol. Soc.*, **142**, 1265–1275.

Schäfler, A., Craig, G., Wernli, H., Arbogast, P., Doyle, J. D., McTaggart-Cowan, R., Methven, J., Rivière, G., Ament, F., Boettcher, M., Bramberger, M., Cazenave, Q., Cotton, R., Crewell, S., Delanoë, J., Dörnbrack, A., Ehrlich, A., Ewald, F., Fix, A., Grams, C. M., Gray, S. L., Grob, H., Groß, S., Hagen, M., Harvey, B., Hirsch, L., Jacob, M., Kölling, T., Konow, H., Lemmerz, C., Lux, O., Magnusson, L., Mayer, B., Mech, M., Moore, R., Pelon, J., Quinting, J., Rahm, S., Rapp, M., Rautenhaus, M., Reitebuch, O., Reynolds, C. A., Sodemann, H., Spengler, T., Vaughan, G., Wendisch, M., Wirth, M., Witschas, B., Wolf, K. and Zinner, T. (2018) The North Atlantic Waveguide and Downstream Impact Experiment. *Bull. Am. Meteorol. Soc.*, **99**, 1607–1637.

Selz, T. (2019) Estimating the intrinsic limit of predictability using a stochastic convection scheme. *J. Atmos. Sci.*, **76**, 757–765.

Selz, T. and Craig, G. C. (2015) Upscale error growth in a high-resolution simulation of a summertime weather event over europe. *Mon. Weather Rev.*, **143**, 813–827.

Stoelinga, M. T. (1996) A potential vorticity-based study of the role of diabatic heating and friction in a numerically simulated baroclinic cyclone. *Mon. Weather Rev.*, **124**, 849–874.

Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P. L., Wilkinson, J., Willett, M., Woolnough, S. and Xavier, P. (2017) The Met Office Unified Model Global Atmosphere 6.0/6.1 and JULES Global Land 6.0/6.1 configurations. *Geosci. Model Dev.*, **10**, 1487–1520.

Whitehead, J. P., Jablonowski, C., Kent, J. and Rood, R. B. (2015) Potential vorticity: measuring consistency between GCM dynamical cores and tracer advection schemes. *Q. J. R. Meteorol. Soc.*, **141**, 739–751.

Wilks, D. S. (2011) *Statistical methods in the atmospheric sciences*. Amsterdam: Academic Press, Elsevier, 3rd edn.

Wood, N., Staniforth, A., White, A., Allen, T., Diamantakis, M., Gross, M., Melvin, T., Smith, C., Vosper, S., Zerroukat, M. and Thuburn, J. (2014) An inherently mass-conserving semi-implicit semi-lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Q. J. R. Meteorol. Soc.*, **140**, 1505–1520.

Zhang, F., Bei, N., Rotunno, R., Snyder, C. and Epifanio, C. C. (2007) Mesoscale predictability of moist baroclinic waves: convection-permitting experiments and multistage error growth dynamics. *J. Atmos. Sci.*, **64**, 3579–3594.

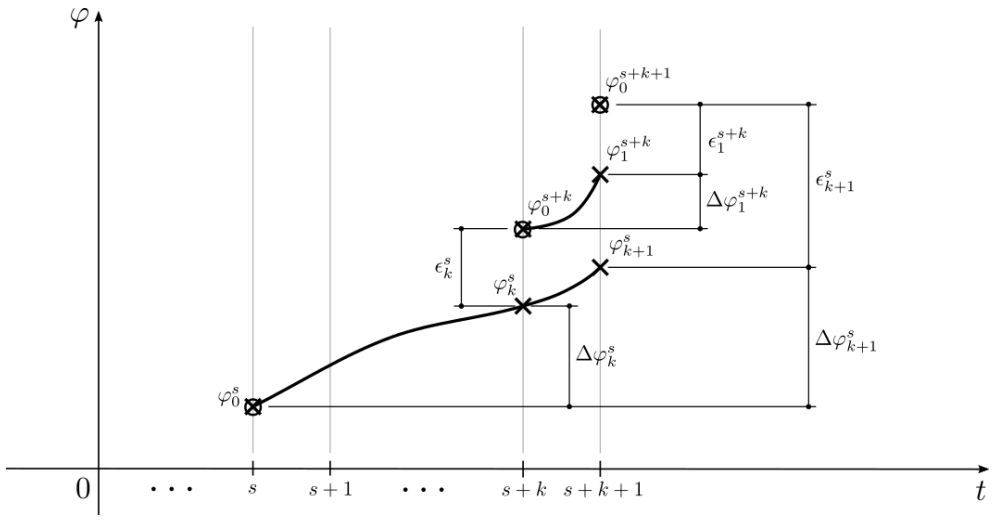

FIGURE 1 Schematic illustrating the relationships between the terms in the Eulerian description at a given grid point. Crosses and circled crosses represent forecasts and analyses, respectively.

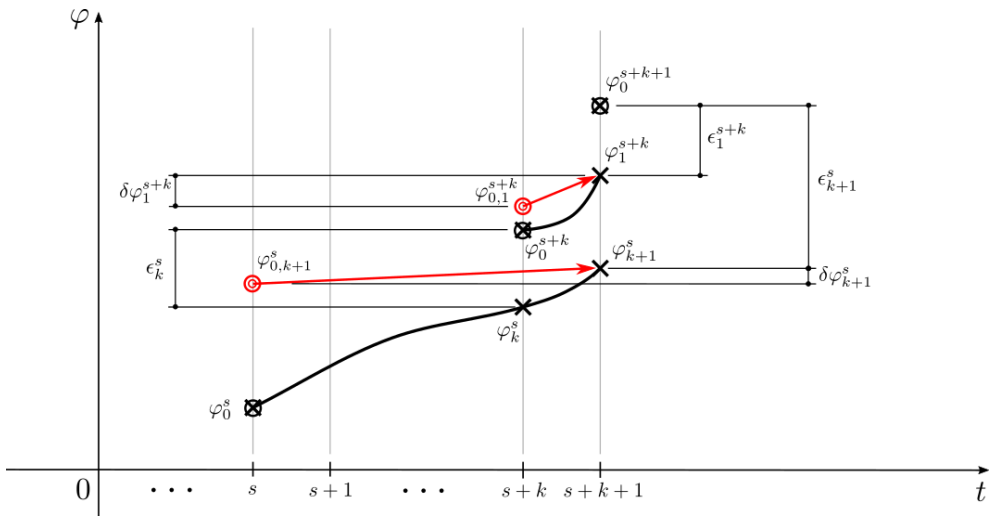

FIGURE 2 Schematic illustrating the relationships between the terms in the Lagrangian-tracer description. Crosses and circled crosses represent forecasts and analyses, respectively. Concentric circles represent the φ-value, at analysis time, belonging to parcels which will be advected to the grid points of interest in the forecasts. The red lines represent the evolutions of these parcels in the forecasts.

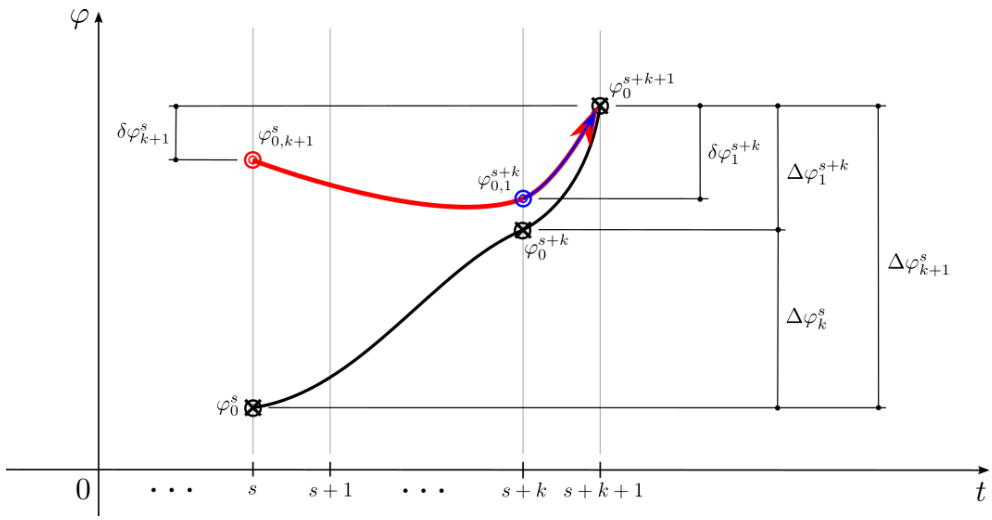

FIGURE 3 Schematic illustrating the behaviour of a perfect model with perfect initial conditions. Circled crosses represent analyses. Concentric circles represent the φ-value, at analysis time, belonging to the parcel which will be advected to the grid point of interest in the forecast. The red and blue lines represent the evolutions of these parcels in the long and short forecast, respectively.

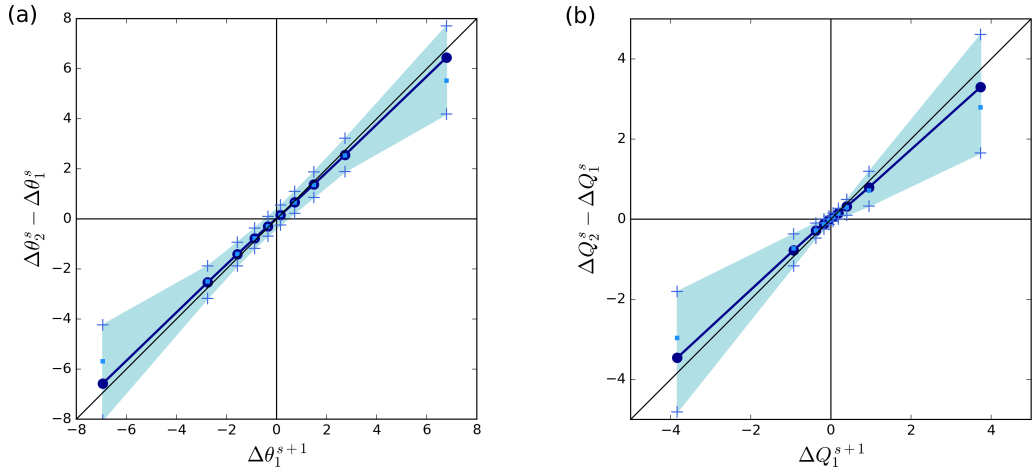

FIGURE 4 $\langle \Delta\varphi_{k+1}^s - \Delta\varphi_k^s \rangle$ versus $\langle \Delta\varphi_1^{s+k} \rangle$ (black circles) for (a) $\varphi = \theta$ (in K) and (b) $\varphi = Q$ (in PVU) for $k = 1$ within ten equally populated bands according to $\Delta\varphi_1^{s+k}$. Small squares represent the median, and crosses represent the 1st and 3rd quartiles of $(\Delta\varphi_{k+1}^s - \Delta\varphi_k^s)$ in each of those bands. The data points are plotted at the position of $\langle \Delta\varphi_1^{s+k} \rangle$ within the corresponding band. For emphasis, the black line joins the means and the light blue shading highlights the position of the interquartile range.

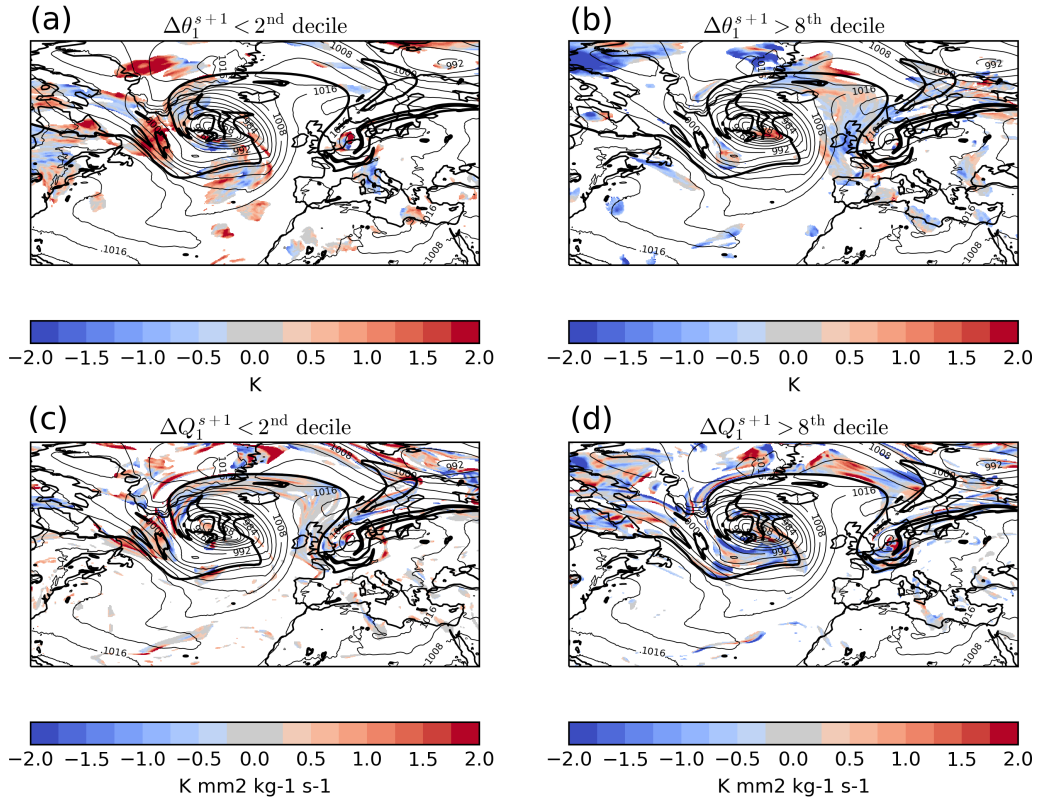

FIGURE 5 Grid points for which $\Delta\varphi_1^{s+k}$ is (a,c) less than its second decile, and (b,c) greater than its eighth decile for (a,b) $\varphi = \theta$ and (c,d) $\varphi = Q$. The grid points in the figure correspond to model level 35 (~ 8.6 km) for $k = 1$ for a single long forecast with base time at 1200 UTC 1 October 2016 ($s = 23$) shaded by $(\Delta\varphi_{k+1}^{s} - \Delta\varphi_k^s) - \Delta\varphi_1^{s+k}$. Thin lines represent mean sea level pressure contours, in hPa, with a separation of 4 hPa; bold lines represent the 320-K 2-PVU PV contour. Both sets of contours correspond to validation time 1200 UTC 2 October 2016.

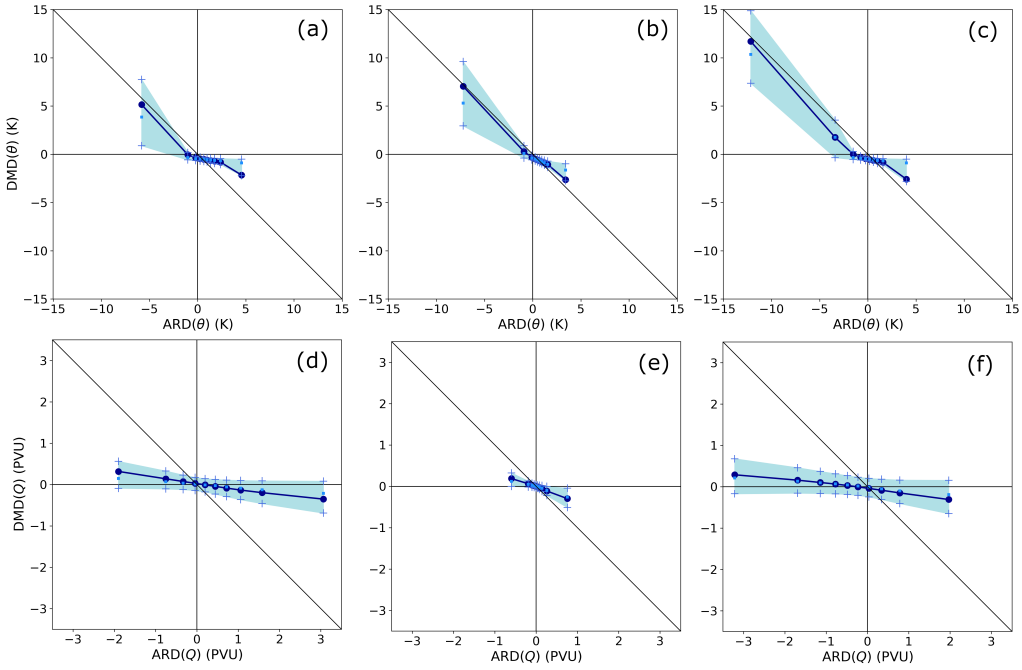

FIGURE 6 $\langle \mathrm{DMD}(\varphi) \rangle$ versus $\langle \mathrm{ARD}(\varphi) \rangle$ (circles) for (a,b,c) $\varphi = \theta$ (in K) and (d,e,f) $\varphi = Q$ (in PVU) for $k = 1$ within ten equally populated bands according to $\mathrm{ARD}(\varphi)$ for grid points for which $\Delta\varphi_1^{s+k}$ is (a,d) below its first decile (b,e) between its fourth and sixth deciles, and (c,f) above its ninth decile. Small squares represent the median, and crosses represent the first and third quartiles of $\mathrm{DMD}(\varphi)$ in each band. The data points are plotted at the position of $\langle \mathrm{ARD}(\varphi) \rangle$ within the corresponding band. The black line with slope −1 passing through the origin describes the expected mean behaviour of an unbiased forecast model. For emphasis, the black line joins the means and the light blue shading highlights the position of the interquartile range.

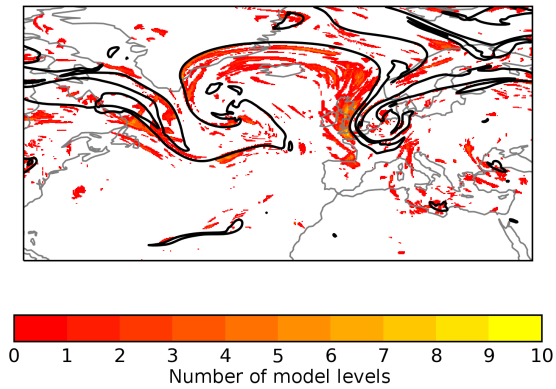

FIGURE 7 Number of tropospheric grid points in each model column, out of ten model levels ($31 \leq m_l \leq 40$, nominally between 6.8 km and 11.2 km), for which $|\Delta\theta_1^{s+k}|$ and $|\Delta Q_1^{s+k}|$ are greater than their respective sixth decile and ARD(Q) is greater than its eighth decile. The grid points in the figure correspond to $k = 1$ for a single 24-h forecast with base time at 1200 UTC 1 October 2016 ($s = 23$). Bold lines represent the 320-K 2-PVU PV contour at validation time 1200 UTC 2 October 2016.