RESEARCH ARTICLE

Dependence on initial conditions versus model formulations for medium-range forecast error variations

Linus Magnusson[1] | Jan-Huey Chen[2,3] | Shian-Jiann Lin[2] | Linjiong Zhou[2,4] | Xi Chen[2,4]

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

[2]National Oceanic and Atmospheric Administration/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey, USA

[3]University Corporation for Atmospheric Research, Boulder, Colorado, USA

[4]Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey, USA

Correspondence

Linus Magnusson, ECMWF, Shinfield Park, RG29AX, Reading, UK.
Email: linus.magnusson@ecmwf.int

Abstract

Understanding the root causes of forecast errors and occasional very poor forecasts is essential but difficult. In this paper we investigate the relative importance of initial conditions and model formulation for medium-range errors in 500 hPa geopotential height. The question is addressed by comparing forecasts produced with ECMWF-IFS and NCEP-GFS forecasting systems, and with the GFDL-fvGFS model initialized with the ECMWF and NCEP initial conditions. This gives two pairs of configurations that use the same initial conditions but different models, and one pair with the same model but different initial conditions. The first conclusion is that the initial conditions play the major role in differences between the configurations in terms of the average root-mean-square error for both Northern and Southern Hemispheres as well as Europe and the contiguous US (CONUS), while the model dominates the systematic errors. A similar conclusion is also found by verifying precipitation over low latitudes and the CONUS. The day-to-day variations of 500 hPa geopotential height scores are exemplified by one case of a forecast bust over Europe, where the error is found to be dominated by initial errors. The results are generalized by calculating correlations between errors integrated over Europe, CONUS and a region in the southeastern Pacific from the different configurations. For Europe and southeast Pacific, the correlations in the medium range are highest between the pairs that use the same initial conditions, while over CONUS they are highest for the pair with the same model. This suggests different mechanisms behind the day-to-day variability of the score for these regions. Over CONUS the link is made to the propagation of troughs over the Rockies, and the result suggests that the large differences in parametrizations of orographic drag between the models play a role.

KEYWORDS

forecast busts, forecast skill, numerical weather prediction

1 | INTRODUCTION

Medium-range weather forecasts have undergone significant improvements since the advent of global forecasts 40 years ago (Bauer *et al.*, 2015). The improvement in skill comes from improved data assimilation, model dynamics and physics, model resolution (horizontal, vertical and temporal), representation of uncertainties, together with improved observing

systems such as satellites, and the usage of these observations. The relative contributions to forecast improvement over the past decades from initial conditions and forecast model were investigated by Magnusson and Källén (2013). They found that similar contributions to the improvements from these two aspects have been seen over the past decades. The fastest pace of improvement for 500 hPa geopotential height (z500) forecasts from the European Centre for Medium-range Weather Forecasts (ECMWF) were between 1995 and 2005 with the introduction of three-dimensional and subsequentially four-dimensional variational data assimilation in 1997 (Rabier *et al.*, 2000) and following improvements in the usage of satellite data. However, the quality of the initial conditions is also highly dependent on the model errors through the use of a short forecast as a "first guess" for the data assimilation and the use of the model within the data assimilation system.

The predictive skill of a forecast is highly flow-dependent in midlatitudes and shows large variations from day to day. Occasionally, forecasts experience very low scores and such episodes are often referred to as "forecast busts" or "dropouts". The mechanisms behind such busts in medium-range forecasts have been investigated by, for example, Rodwell *et al.* (2013) and Magnusson (2017). Even though it is clear that the atmospheric dynamics play the major role in the error amplification, it is still unclear if the major source of the error originates from the initial conditions or appears during the model integration because of errors in the model formulation and/or computational errors.

Several different techniques to track the origin of forecast errors are available. Magnusson (2017) discussed the usefulness of manual error tracking, ensemble sensitivities based on work from Torn and Hakim (2008) and nudging experiments. These techniques indicate the geographical origin of the error, but cannot disentangle errors in initial conditions from those arising during the forecast integration due to model/computational errors. One method to address this question is to use the same initial conditions in different models and compare with forecasts using different initial conditions in the same model. However, this method requires a global model which can be initialized from different analyses and usually needs major collaborations between different modelling centres, e.g. as in the ensemble evaluation by Richardson (2001) and Harrison *et al.* (1999). In Rodwell *et al.* (2012), the UK Met Office (UKMO) model was used and reproduced a case of a poor ECMWF forecast over Europe in April 2011 using the same ECMWF analysis as initial conditions.

In this study, based on a collaboration between ECMWF and National Oceanic and Atmospheric Administration Geophysical Fluid Dynamics Laboratory (NOAA/GFDL), the functionality to use ECMWF analyses as the model initial conditions was built into the GFDL finite-volume Global

Forecasting System (fvGFS; Chen *et al.* 2018a) to evaluate the forecast difference from using the National Centers for Environmental Prediction (NCEP) GFS and ECMWF analyses as the initial conditions. The model fvGFS has been developed and maintained at GFDL. This experimental set-up gives two pairs of forecasts using the same initial conditions but different models, and one pair with the same model but different initial conditions. In this paper we will evaluate these forecasts in midlatitudes in order to address the question of the dependency of model versus initial errors with respect to the day-to-day variability in z500 scores. It should be noted that the initial conditions are partly a product of the model used in the data assimilation system, i.e. the initial conditions are a product of the observation usage, data assimilation method and the model. However, if a difference appears between forecasts starting from the same initial conditions but different models, it will highlight dependency on the model formulation. Even if we were not able to fully answer the question of the impact from the model formulation for the full forecasting system, the results for this study are important, for example to guide development of ensemble systems and to interpret differences between forecasts from different forecasting systems.

We are focusing here on z500 to understand the impact on the synoptic predictability. Medium-range forecast errors in the surface parameters such as 2 m temperature are often dependent on the ability to predict the large-scale flow, such as in the case of the European forecast bust presented in Grams *et al.* (2018).

The structure of the paper is as follows. In Section 2 the different models and data assimilation systems are briefly described. In Section 3 the results are presented for the scores for a sample covering a full year, for one case of a forecast bust in the ECMWF forecast and for the correlation of the day-to-day variations in the full sample. In Section 4 the results from the previous section are further explored, and finally the results are summarized in Section 5.

2 | MODEL AND INITIAL CONDITION DESCRIPTIONS

2.1 | ECMWF IFS

The ECMWF operational forecasts are produced with the ECMWF Integrated Forecasting System (IFS). The section below gives a brief description of the model system and details can be found in ECMWF (2016). The dynamical core of the model is hydrostatic and spectral, and employs a two-time-level semi-implicit semi-Lagrangian scheme combined with a spectral transform technique for the horizontal discretization and a finite-element method for the vertical discretization (Untch and Hortal, 2004). The current resolution for the ECMWF high-resolution, deterministic forecast

(HRES) is approximately 9 km in the horizontal, and with 137 vertical levels reaching up to 0.01 hPa (approximately 80 km).

The radiation code is based on the Rapid Radiation Transfer Model and cloud–radiation interactions are taken into account using the McICA (Monte Carlo Independent Column Approximation) method (Morcrette *et al.*, 2008). The parametrization of convection is based on the mass-flux approach (Bechtold *et al.*, 2008) with a modified CAPE closure leading to an improved diurnal cycle of convection (Bechtold *et al.*, 2014). The cloud and large-scale precipitation scheme is based on Tiedtke (1993), but has been substantially upgraded with separate prognostic variables for cloud water, cloud ice, rain, snow and cloud fraction, and improved parametrization of microphysical processes (Forbes and Ahlgrimm, 2014). The orographic gravity wave drag is parametrized following Lott and Miller (1997) and Beljaars *et al.* (2004), and a non-orographic gravity wave drag parametrization is described in Orr *et al.* (2010). Recent modifications to these schemes include reduced turbulent mixing in stable conditions and increased orographic drag (Sandu *et al.*, 2014). The surface module of IFS is described in Balsamo *et al.* (2014).

The data assimilation component of IFS consists of a four-dimensional variational data assimilation (4D-Var; Rabier *et al.* (2000)). The initial conditions for the operational forecasts use a 6 hr assimilation window (±3 hr of the initialization time), while the first-guess forecast is provided from an analysis based on a 12 hr window. To provide background-error statistics, a 25-member ensemble of 4D-Var assimilations (Ensemble Data Assimilation, or EDA) is run with a lower horizontal resolution (Bonavita *et al.*, 2012).

The forecasts and initial conditions used in this study are based on model cycle 41r2, which became operational in March 2016, with 9 km resolution and 137 vertical levels. This study uses initial conditions and forecasts from the pre-operational testing up to 8 March 2016, when the cycle was implemented, and operational data thereafter. During the evaluation period studied in this article, the model (HRES) used persisted SST anomalies during the integration.

2.2 | NCEP GFS

The atmospheric forecast model used in the NCEP GFS is a global spectral model (GSM) with spherical harmonic basis functions. Details about the GFS model system can be found in http://www.emc.ncep.noaa.gov/GFS/doc.php (accessed 18 April 2019). The current operational horizontal resolution is approximately 13 km at the Equator for forecast days 0–10. In the vertical there are 64 hybrid sigma-pressure (Sela, 2009) layers with the top layer centred around 0.27 hPa (approximately 55 km). The current operational dynamical core of the GFS is based on a two-time-level semi-implicit semi-Lagrangian discretization (Sela, 2010) with three-dimensional Hermite interpolation.

The radiation parametrizations are modified and optimized versions of the Rapid Radiative Transfer Models (Clough *et al.*, 2005). A hybrid eddy-diffusivity mass-flux (EDMF) PBL parametrization has been used since January 2015. The orographic gravity wave drag and mountain blocking parametrization is scale-aware in the GFS and implemented across NCEP global and regional models (Chun and Baik, 1994; Kim and Arakawa, 1995; Kim and Doyle, 2005). For deep and shallow convection, the Simplified Arakawa–Schubert (SAS) parametrization (Han and Pan, 2011) is used. The Zhao–Carr grid-scale scheme is used for condensation and precipitation parametrization (Zhao and Carr, 1997). The land surface model (LSM) of GFS is four soil layer (10, 30, 60, 100 cm thick) Noah model (Ek *et al.*, 2003).

The initial conditions for the global forecasts are obtained through the Global Data Assimilation System (GDAS) which uses a hybrid four-dimensional ensemble variational formulation (Hybrid 4DEnVar; Buehner *et al.* (2013)). The GDAS ingests all available observations within a ±3 hr window of the analysis time. A 9 hr GSM forecast (TL1534 interpolated to TL574) from the previous GDAS analysis is used as the first guess for the assimilation. The GDAS also runs with a late data cut-off to provide the first-guess forecast for the next 6-hourly cycle.

2.3 | NOAA/GFDL fvGFS

The NOAA/GFDL fvGFS model combines the Finite-Volume Cubed-Sphere Dynamical Core (FV3) and the common GFS physics package, which was initially provided for the Next Generation Global Prediction System (NGGPS) phase II to test the robustness of the dynamical cores under a wide range of realistic atmospheric initial conditions. FV3 uses the "vertically Lagrangian" dynamics of Lin (2004) extended with a non-hydrostatic pressure-gradient computation of Lin (1997) and a semi-implicit solver for vertically propagating sound waves, discretized on the cubed-sphere grid of Putman and Lin (2007). An alternative vertically Lagrangian non-hydrostatic discretization is also demonstrated in Chen *et al.* (2013b). The GFS physical parametrizations described in the previous section are used in this study, except for the Zhao–Carr grid-scale condensation and precipitation parametrization which is replaced with the GFDL single-moment six-category cloud microphysics (Zhou *et al.*, 2019). In the most recent version of fvGFS, the EDMF PBL scheme is replaced with the YSU (Yonsei University) PBL scheme (Hong *et al.*, 2006) and a mixed-layer ocean model (Polland *et al.*, 1973) with some modifications, e.g. including an ocean current damping term, a relaxation of the ocean mixed-layer depth toward observational climatology and a relaxation of the SSTs toward observational climatology plus initial anomaly, is also used. Initialization of the atmosphere, land surface, and sea-surface temperatures (SSTs) are from NCEP operational analyses. A 13 km uniform-resolution

version of fvGFS has been run in real time at GFDL since mid-2016. The forecast characteristics of the fvGFS, with a focus on tropical cyclone prediction, are described in Chen *et al.* (2018a) and Hazelton *et al.* (2018).

Besides the pre-processing tool developed during the NGGPS phase II to use the GFS analysis as the initial condition, a sophisticated interpolation tool was also developed in fvGFS to carefully use the IFS analysis data as the initial condition. The interpolation procedure from the Gaussian grid to the cubed-sphere grid is documented in Chen *et al.* (2018b). All upper-air variables are initialized from the IFS analysis, while the land surface variables are initialized from the GFS analysis for practical reasons. This discrepancy will affect land surface temperatures and also sub-seasonal predictions, but is believed to have a small impact on medium-range upper-air variability which is the scope of this study.

3 | RESULTS

To compare different forecast configurations over a full year, forecasts are initialized/used every fifth day between 15 August 2015 and 9 August 2016. By running forecasts every fifth day, the temporal correlation of errors is expected be small. EC is based on pre-operational runs before 8 March 2016 and operational runs of ECMWF HRES forecast thereafter. GFS is based on operational NCEP deterministic forecasts, while FV3gfs uses the GDFL fvGFS initialized from NCEP initial conditions and FV3ec uses fvGFS initialized from ECMWF initial conditions.

The verification presented in this paper is based on z500 on a 1° regular lat–lon grid. The forecasts are verified against the UKMO analysis, provided by the TIGGE archive (Bougeault *et al.*, 2010), so as not to favour any of the configurations. Mean error (ME), root-mean-square error (RMSE) and anomaly correlation coefficients (ACC) have been calculated and evaluated. The scores have been calculated for the Northern Hemisphere (N.Hem, 20°–90°N), Southern Hemisphere (S. Hem, 20°–90°S), Europe (35°–75°N, 12.5°W–42.5°E) and contiguous US (CONUS, 25°–50°N,125°–65°W). While the results presented in the next sections are based on RMSE, similar conclusions hold for ACC.

As a complement to the z500 evaluation, we have also evaluated 6-hourly precipitation forecasts against TRMM 3B42 version 7, which merges satellite rainfall estimates with gauge data (Chen *et al.*, 2013a). The temporal resolution of TRMM 3B42 is 3-hourly, while the horizontal resolution is 0.25° between 50°S and 50°N. The precipitation is evaluated for a zonal band spanning lower latitudes (40°N–40°S) and for CONUS.

3.1 | Average scores

RMSE (Figure 1) have been calculated for N.Hem, S. Hem, Europe and CONUS for the four forecast datasets based on

all cases. Figure 1b,d,f,h show the differences in RMSE with respect to FV3ec. In these panels the statistical significance of the differences with respect to FV3ec has been calculated with a Student's *t*-test, and lead times where the difference is significantly different (95% confidence level) are indicated with dots.

For lead times up to 2–3 days, the EC forecasts show the lowest errors for all evaluated regions, including being lower than FV3ec. One could speculate whether this difference is due to more similarities in the mean climate between the UKMO analysis and ECMWF forecasts. For the 5–7-day forecasts, the scores over N.Hem and S.Hem for EC and FV3ec are similar without any significant differences, and both are significantly better than FV3gfs and GFS. For N.Hem, FV3gfs has lower RMSE than GFS, signalling that the model difference plays a role for the forecast error but such a difference is not prominent for S.Hem. For the smaller regions (Europe, CONUS), the scores are more noisy but gives a similar ordering as for the full Northern Hemisphere.

For the mean error (ME) of z500 (Figure 2), all forecasts initially have a negative bias compared to the UKMO analysis for both hemispheres. This difference is due to a bias in the UKMO data assimilation which was corrected during spring 2016 and hence affected the first part of the sample. FV3ec has somewhat lower geopotential already after 6 hr, although initialized from the same analysis as EC. Later both FV3 systems drift to a significant lower geopotential in N.Hem, while EC has a slight positive drift and GFS shows a slow negative drift.

To further evaluate the average differences between the simulations, Figures 3 and 4 shows the RMSE and ME in precipitation respectively for 40°N–40°S and for the CONUS. Europe has not been included due to a large part being outside of the TRMM region. For 40°N–40°S, all forecasts overestimate the precipitation compared to TRMM, with the largest biases for GFS. As FV3ec is initialized from an analysis from a different model, we can expect some initialization shock even if we apply the initialization procedure outlined in Chen *et al.* (2018b). For the first 6 hr we indeed find the highest bias in FV3ec, but later the bias quickly stabilizes and is similar to FV3gfs. For EC we find a strong diurnal cycle in the precipitation bias for both regions. This is not necessarily due to a real diurnal cycle bias as the verification for 40°N–40°S includes all longitudes, but rather can be a sign of a geographical bias. One contributor here is an over-prediction of precipitation over tropical South America (not shown) in the EC, which contributes to 1200–1800 UTC verification. Evaluating the biases over CONUS, the EC has a strong negative bias for 0000–0600 UTC due to underestimation of evening precipitation over the central US (not shown), but shows small biases for the other times of the day. The diurnal cycle in the ECMWF forecasts were greatly improved in 2013 with an update of the convection scheme discussed in Bechtold *et al.*
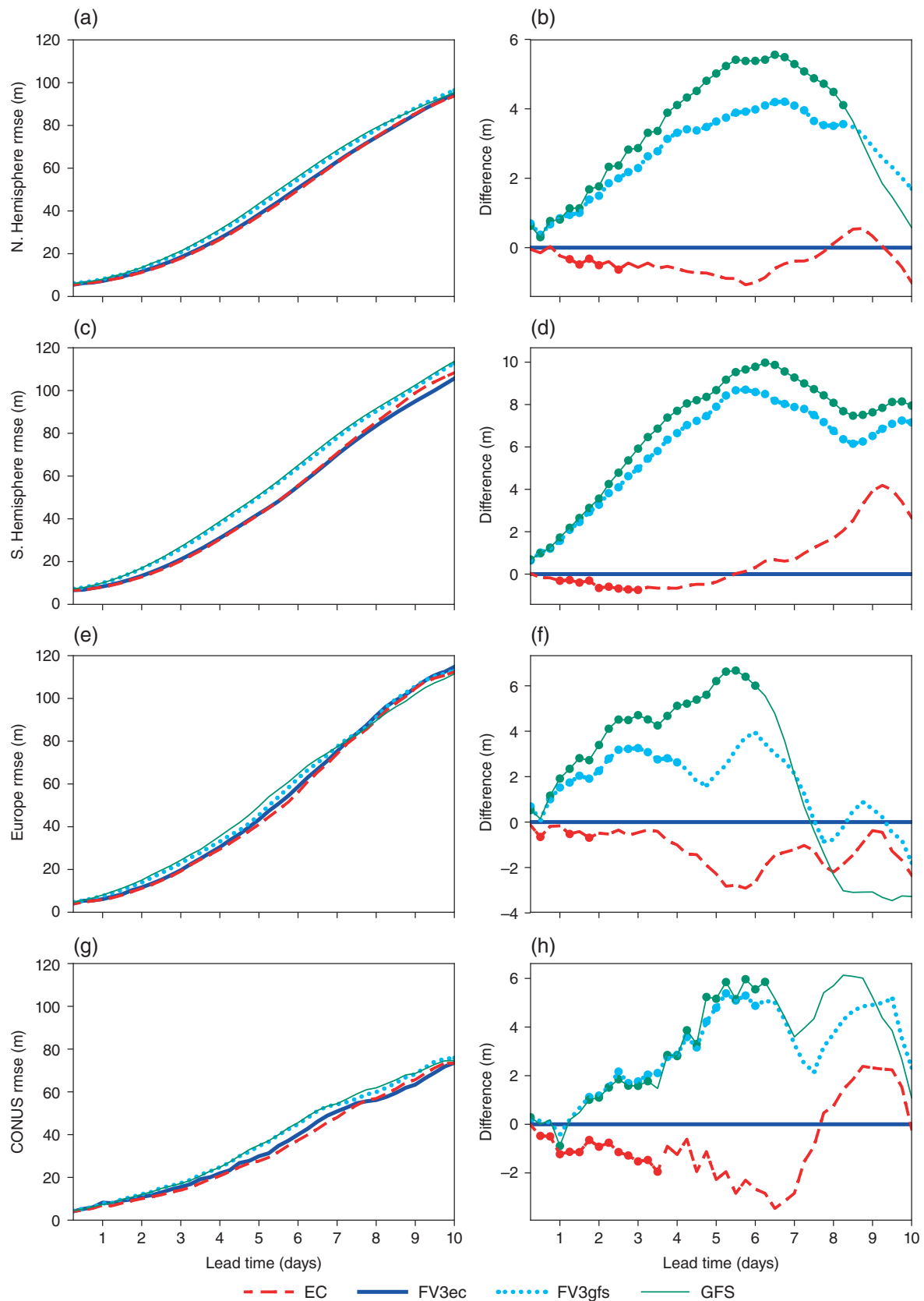
FIGURE 1   (a, c, e, g) RMSE and (b, d, f, h) difference with respect to FV3ec (right) for z500 for (a, b) N.Hem, (c, d) S.Hem, (e, f) Europe and (g, h) CONUS. Thick dots indicate statistically different results with respect to FV3ec
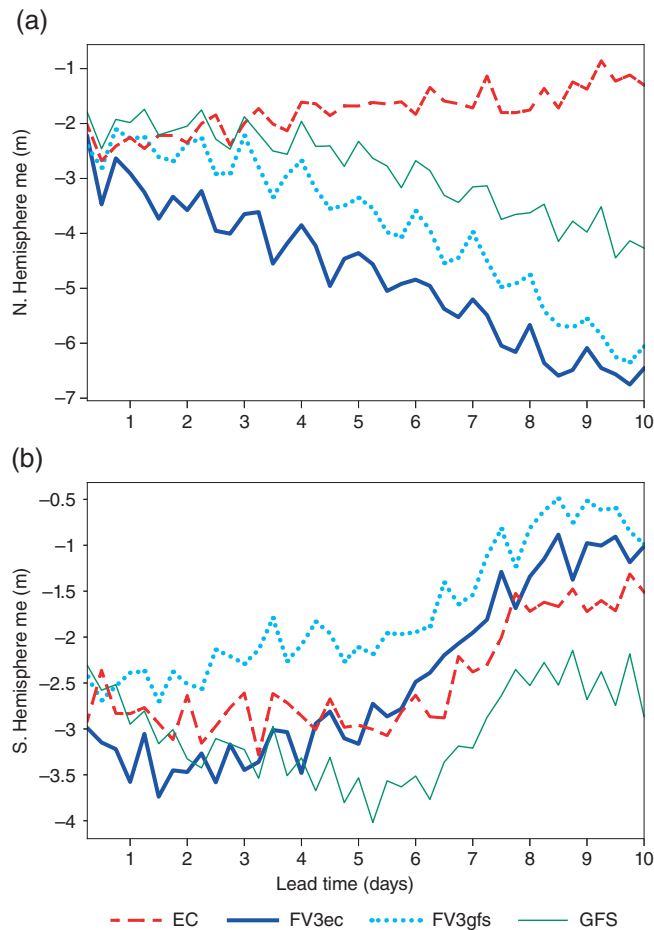
(a)

(b)

FIGURE 2 Mean error versus lead time for z500 for (a) N.Hem and (b) S.Hem

(2014). Both FV3 forecasts and GFS have a more constant negative bias for all times of the day for CONUS.

Regarding the RMSE for precipitation (Figure 3) for 40°N–40°S, we find the lowest errors for all lead times apart from the first day for EC, followed by FV3ec. For the first day, FV3ec obtains the lowest RMSE despite the spin-up issue mentioned in the previous paragraph. The results show that, also for RMSE in precipitation, the initial conditions play an important role. For CONUS, the results are more noisy but also here we find a lower RMSE for the forecasts initialized with ECMWF analyses for the first few days.

The results in this section show that the mid-tropospheric forecast errors in the range of 5–7 days have larger sensitivity to the initial conditions than to the forecast model, with some advantage also for the FV3 model compared to the GFS. For the mean error, the model dependence is much stronger but, despite very different bias structure between EC and FV3ec, their RMSE is similar and superior to that of the forecasts initialized from GFS analyses. In the coming sections we will investigate whether the sensitivity to the choice of initial conditions also holds for the day-to-day variability of the z500 errors.

3.2 | Case-study

To exemplify the errors in the different forecasts, we have chosen one case of a forecast bust for Europe. The case has previously been evaluated in Magnusson (2017) and Grams *et al.* (2018) and we refer to these papers for details of the predictability and error sources. The bust was associated with a failure in predicting a blocking over Europe and resulted in large 2 m temperature errors for northwestern Europe. The case presented here was the most extreme case in terms of ECMWF forecast error during the evaluated period for day 6 over Europe.

Figure 5 shows the time series of RMSE for 6-day z500 forecasts over Europe for EC, GFS, FV3gfs and FV3ec. To give an estimate of the uncertainty in the forecasts, the scores from individual ensemble members are also included from the ECMWF and NCEP operational ensembles, available through the TIGGE archive (Bougeault *et al.*, 2010). The ensemble results are visualized with the median and the area covered by the 10th to 90th percentile of the ensemble members (80% of the members). The figure also includes the ensemble standard deviation (spread) for the both ensembles. Note that the median value of the scores for individual members is not the same as the score of an ensemble mean, and that one should expect on average a lower RMSE from a deterministic unperturbed forecast than from individual ensemble members.

From the time series in Figure 5, variations in scores from day to day are evident due to periods of higher and lower atmospheric predictability. We also find variations of the ranking between the different forecast systems. Although EC and FV3ec on average have lower RMSE for both regions for the full period (Figure 1) than GFS and FV3gfs, this does not hold true for every day, as demonstrated by the case below.

In the beginning of March 2016, both NCEP and ECMWF forecasts experienced increased errors over Europe; this is also apparent considering the errors in both ensembles and we also find an increased ensemble standard deviation. However, the NCEP forecast recovered from the error a day earlier than ECMWF, and for the 7 March 0000 UTC forecast the 50% (median) of the ECMWF ensemble members was worse than 90% of the NCEP ensemble members, and the EC deterministic forecast (red dot) was far worse than GFS (green dot). Comparing the two FV3 forecasts, the FV3ec performed worse than FV3gfs. Magnusson (2017) traced the ECMWF forecast error back to the development of a surface low over the western Atlantic. The error later amplified and propagated eastward in connection with a warm conveyor belt that was triggered ahead of the low (Grams *et al.*, 2018). In the 4–6 day range, the error affected the ridge-building over the northeastern Atlantic and later the onset of a Scandinavian blocking regime.
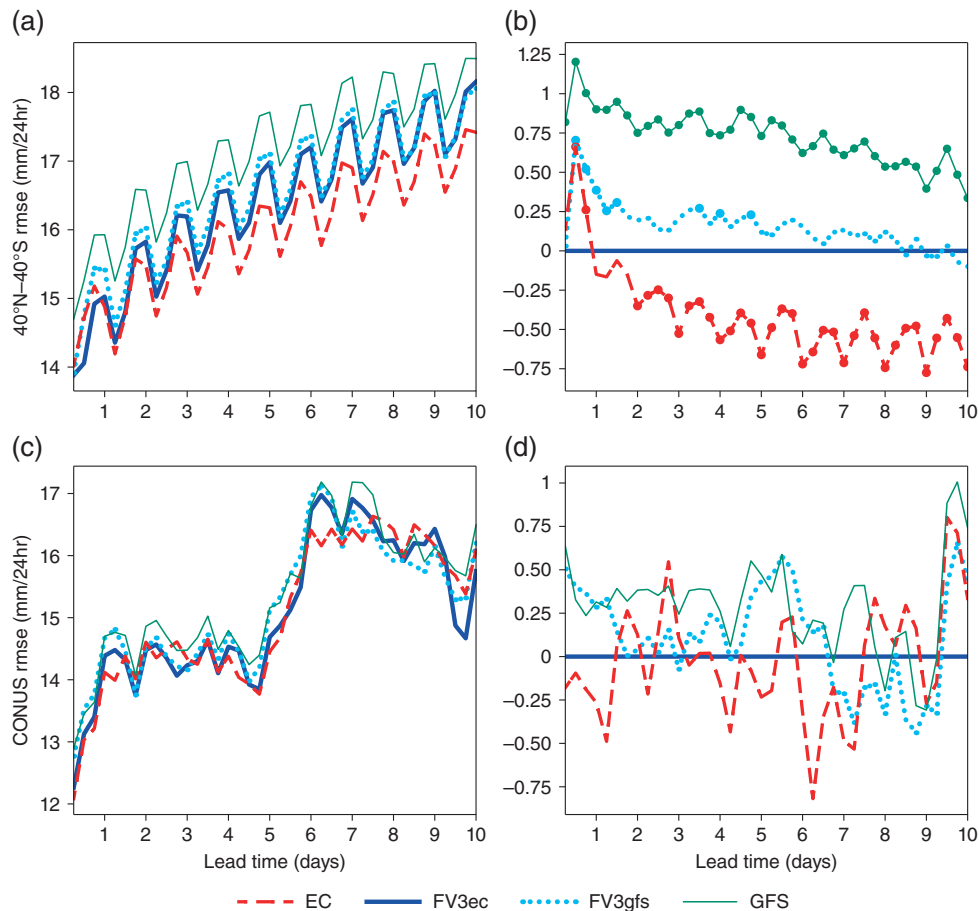
FIGURE 3 (a, c) RMSE and (b, d) difference with respect to FV3ec for 24 hr precipitation. Thick dots in (b) indicate statistically different results with respect to FV3ec. The scores are calculated for (a, b) 40°N–40°S and (c, d) CONUS

Figure 6 illustrates the z500 error in 2-day and 6-day forecasts initialized at 0000 UTC on 7 March from the four different forecast configurations. In the 2-day forecasts, both forecasts initialized from ECMWF initial conditions (EC and FV3ec, a–d) show similar structures of the errors over the central Atlantic, with a too zonal flow compared to the corresponding analysis. Such error structure is not present in the two forecasts initialized with GFS initial conditions (GFS and FV3gfs). Comparing the 6-day error in the different forecasts, the error related to the ridge over northern Europe is much larger in the forecasts with ECMWF initial conditions than in the forecasts with GFS initial conditions. This result indicates that, for this case, errors in the initial conditions played the major role in creating the forecast bust over Europe. In the next section we will try to generalize the results by using the full experiment period.

3.3 | Correlations of errors

To further investigate the relative importance of the initial conditions and model formulations for the day-to-day variations in scores, in this section we examine correlations of errors between the forecast configurations.

Figure 7 shows the Pearson correlation of RMSE versus lead time from different pairs of configurations. A confidence interval (shaded) is included for FV3ec versus FV3gfs based on a bootstrap method where we have generated 1,000 series with the same length (i.e. with replacement) as the original series by randomly sampling from the original series. The confidence interval shows the 5th to 95th percentile of the distribution from the random series. To test the sensitivity for the correlation calculation, the Spearman correlation (not shown) has also been calculated and the results give a message similar to the Pearson correlation.

The reason for showing the sub-regions (Europe, CONUS and SE Pacific in this case) instead of the full hemispheres is that we are interested in the flow-dependent variations of the errors, which are often smoothed out when averaging over the full hemisphere. The region over the SE Pacific is defined as 60°–25°S,120°–75°W. The three regions are outlined in Figure 8.

First of all, we find relatively high correlations in all pairs of experiments for all regions. This correlation is due to the flow-dependence of the error, meaning that forecasts with different model and initial conditions still have correlated errors.
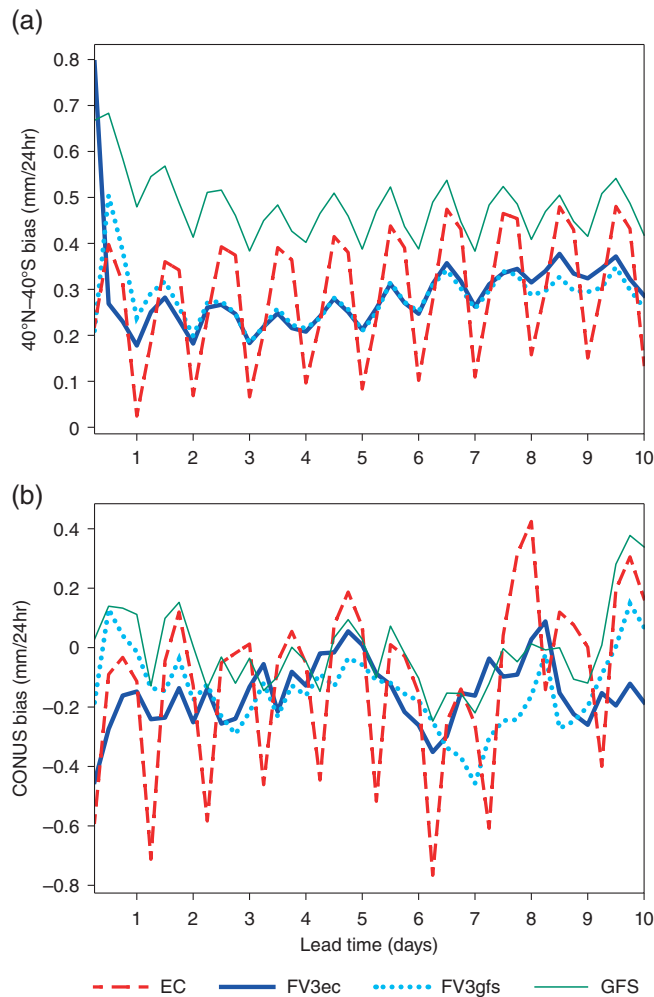
(a)

(b)

FIGURE 4 Mean error for 24 hr precipitation for (a) 40°N–40°S and (b) CONUS

For the shortest lead times, the correlations are highest for the combinations that use the same initial conditions, as expected.

Already at 1-day lead time, differences appear between the three regions. For Europe and SE Pacific, the initial separation between the pairs persists and highest correlations are found for the pairs that share initial conditions (FV3gfs/GFS and FV3ec/EC) for all lead times well into the medium range. On other hand, the correlation between FV3ec/FV3gfs error is lower and on a similar level to the combinations that do not share model or initial conditions (FV3ec/GFS, FV3gfs/EC and GFS/EC). This indicates that also for the full period, the initial conditions play the largest role for the error variations in z500 over Europe and SE Pacific for all time-scales, including the medium range.

As for Europe, for CONUS the highest correlation is found for FV3gfs/GFS. We need to keep in mind that FV3gfs and GFS share both the initial conditions and many of the physical parametrizations in the model. In contrast to the results for Europe, the correlations for FV3ec/EC are lower than for FV3gfs/FV3ec for days 1–3 and after that are on a similar level. The correlation for days 1–3 for FV3ec/EC is also

slightly lower than for FV3ec/GFS, where the latter pair shares large parts of the model physics. This result indicates that the physical parametrizations might play a larger role in the day-to-day variation of the errors for North America than for Europe, especially for relatively short-range forecasts. In the next section we discuss these results.

4 | DISCUSSION

To understand the regional differences in the error correlations found in the previous section, Figure 8 shows the RMS difference (RMSD) in z500 between (a, b) FV3gfs and GFS, (c, d) FV3ec and EC, and (e, f) FV3ec and FV3gfs after 6 hr and 30 hr into the forecasts. The panels include the outlines of the regional boxes used in Figure 7 plus a smaller box over central CONUS. The 6 hr RMSD should be close to the initial RMSD and is expected to be small for FV3gfs/GFS, and FV3ec/EC. However, the initialization procedure could create some initial differences, and also there may be rapidly developing differences due to model formulations. For the FV3ec/EC pair we find the largest 6 hr RMSD where we have the highest orography.

The RMSD between FV3ec and FV3gfs should reflect the differences between the ECMWF and GFS analyses. Here the largest RMSD are found in the Southern Hemisphere, where the observation networks are more sparse and hence the analyses are less constrained. In the Northern Hemisphere the largest RMSD are found for a similar reason over the Arctic and over oceans. Regarding the RMSD between FV3ec and FV3gfs after 30 hr, we find the largest values over the southern oceans and also over the Atlantic. The region with large RMSD over the Atlantic partly falls into the Europe box.

As synoptic-scale errors (and forecast differences) in the first part of medium-range forecasts grow roughly exponentially (Lorenz, 1982; Magnusson and Källén, 2013), forecasts with large initial differences will separate faster than forecasts that start with very small initial differences. Such separation is faster in regions with strong baroclinicity. For the Southern Hemisphere, the relatively large difference in initial conditions is a likely explanation of the low error correlations between the forecasts with different initial conditions compared to the ones that share initial conditions. This is also likely to be the explanation for the dependency on the initial conditions for the day-to-day variability of score over Europe, as it is downstream of the Atlantic with relatively large initial differences and a dynamically active region with fast baroclinic growth of disturbances (not shown).

However, the argument above should also apply for CONUS which is downstream of the Pacific, while we here find higher error correlations for the forecast with the same model. As the shift in the ordering appears during the first
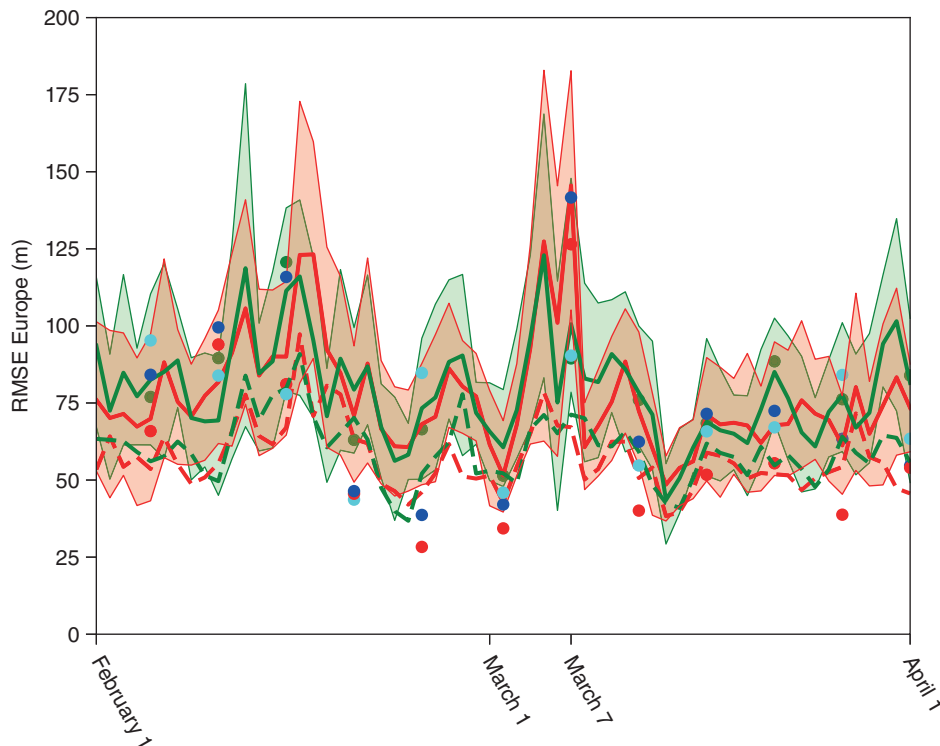
FIGURE 5 Time series from 1 February to 1 April 2016 of RMSE for z500 verified over Europe for day 6 for EC (red dot), GFS (green dot), FV3ec (blue dot) and FV3gfs (cyan dot). Bold lines show the ensemble median of RMSE and shading indicates the range covered by 80% of the ECMWF (red) and NCEP (green) ensemble members. Dashed lines denote the ensemble standard deviation

18 hr of the forecasts (Figure 7b), we can expect local processes to play a key role. It is also worth noting that the correlation after 6 hr is still largest for the pairs that share initial conditions, and is therefore not simply a product of the initialization process in FV3 from ECMWF initial conditions.

After 30 hr, we find a local maximum of RMSD between EC and FV3ec over central North America (Figure 8) on the eastern side of the Rockies. The magnitude is similar to (and locally higher than) what we find between FV3ec and FV3gfs.

To understand the local maximum in the RMSD between EC and FV3ec over central CONUS, we make a composite of the seven cases (10% of the sample) with the largest growth of RMSD (dRMSD) between 6 and 30 hr. The selection is based on dRMSD in the region of 30°–50°N, 110°–90°W (outlined in Figure 9). To inspect the evolution of mean z500 for these cases, the composite of the analyses from 6 hr before, 18 hr after, and 42 hr after the forecast dates are plotted in Figure 9. The middle point ("18 hr after") is chosen to be in the middle of the window used for the difference growth. The difference from zero anomaly has been tested with a Student t-test, and areas that pass the 95% confidence limit are plotted in bold colours.

In the composite of the analyses 6 hr before the initialization, a ridge was present over the eastern Pacific and a trough was centred over the west coast of North America. At 24 hr later, the trough was passing the Rockies and starting to

amplify in the lee of the mountains. Another day later (42 hr after the initialization time of the forecasts), the trough over central US had further amplified and we also find a ridge further downstream forming part of a Rossby wave packet. The location of the trough agrees with the region of the large RMSD between FV3ec and EC after 30 hr.

These composites suggest that the cases of most rapid dRMSD between FV3ec and EC are related to the passage of troughs over the Rockies. As the formation of a trough in the lee of mountains is dependent on the conservation of potential vorticity (e.g. Holton 1993), differences in diabatic processes related to orography between FV3 and EC seem to be possible causes for the growth in z500 difference.

To investigate the difference between FV3ec and EC related to orography, Figure 10 shows diagnostics of surface drag from the parametrization of the sub-grid turbulence and gravity waves plus blocking (ECMWF, 2016) . Comparing the two components, we find large differences in the partitioning between the surface drag processes in the two models, where the main part of the drag in the IFS model is from the sub-grid turbulence, while in the FV3 the main drag over orography is from the gravity-wave part. The sum of the components over the Rockies results in 5% more drag in the FV3 model. It is also worth noting that the spatial pattern of the drag is different in the two models, with more localized drag over high mountains in FV3. The difference could be due to the
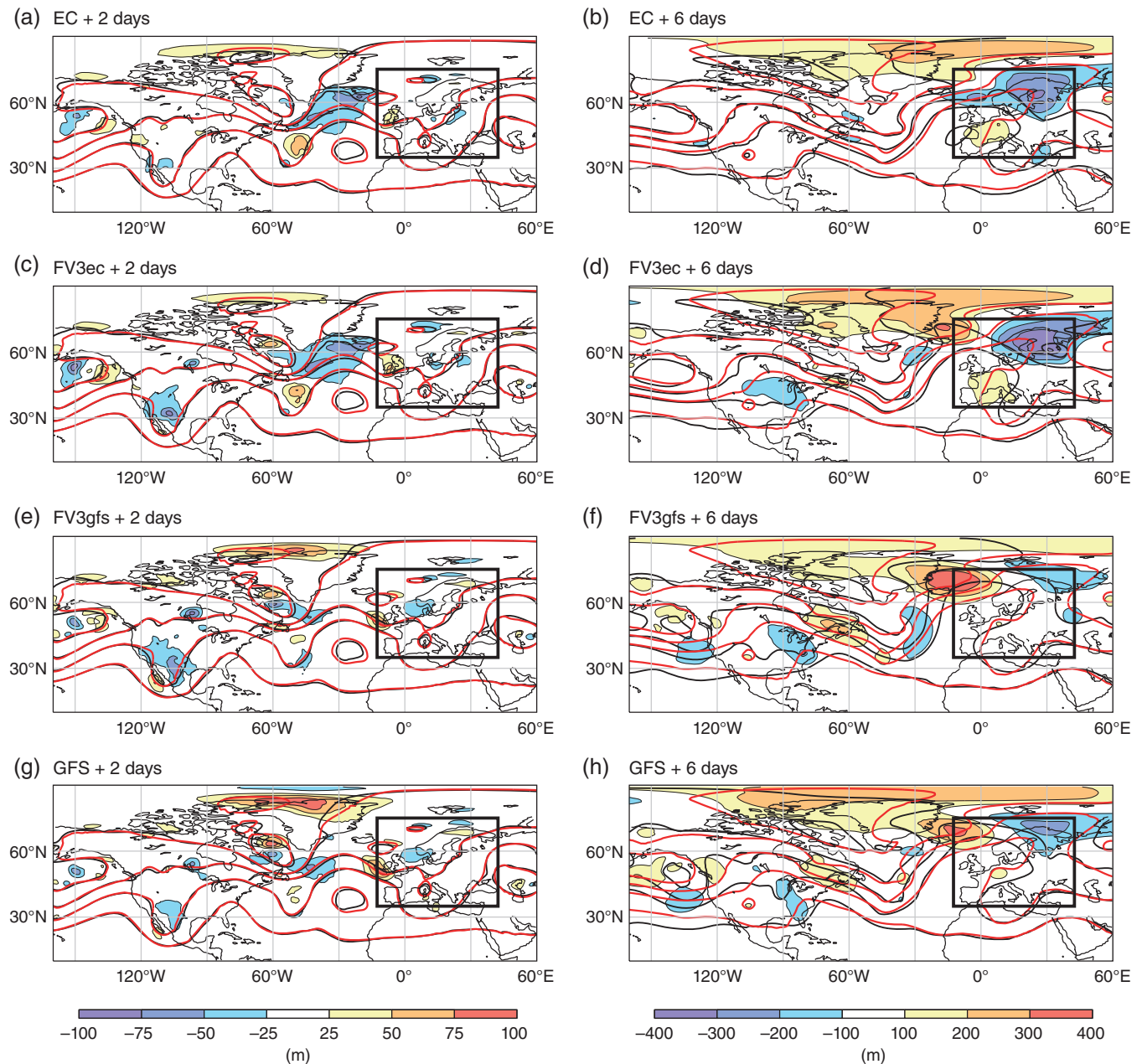
FIGURE 6 z500 forecasts (black contours) and corresponding analyses (red contours) and error (m, shading) in (a, c, e, g) 2-day and (b, d, f, h) 6-day forecasts initialized at 0000 UTC on 7 March. Forecasts are from (a,b) EC, (c,d) FV3ec, (e,f) FV3gfs and (g,h) GFS

parametrizations of the orographic drag and/or the definition of the orography in the different models. As the orographic drag has a day-to-day variability due to the large-scale flow, it could therefore result in differences between the two models and at least partly explain the day-to-day variability in the scores.

As discussed in the previous section, there are also significant differences in the diurnal precipitation cycle between EC and both FV3 systems over CONUS that could also contribute to the differences in z500 for the simulations starting from the same initial conditions, but interacting with the large-scale flow on slightly different time-scales. However, we did not

find any strong correlation between convective precipitation and cases with fast dRMSD between FV3ec and EC (not shown).

5 | SUMMARY AND CONCLUSIONS

In this paper we have investigated the relative importance of initial conditions and model formulation for errors in 500 hPa geopotential height (z500), both in terms of average root-mean-square error (RMSE) and mean error as well as the day-to-day variability, and average scores for precipitation.
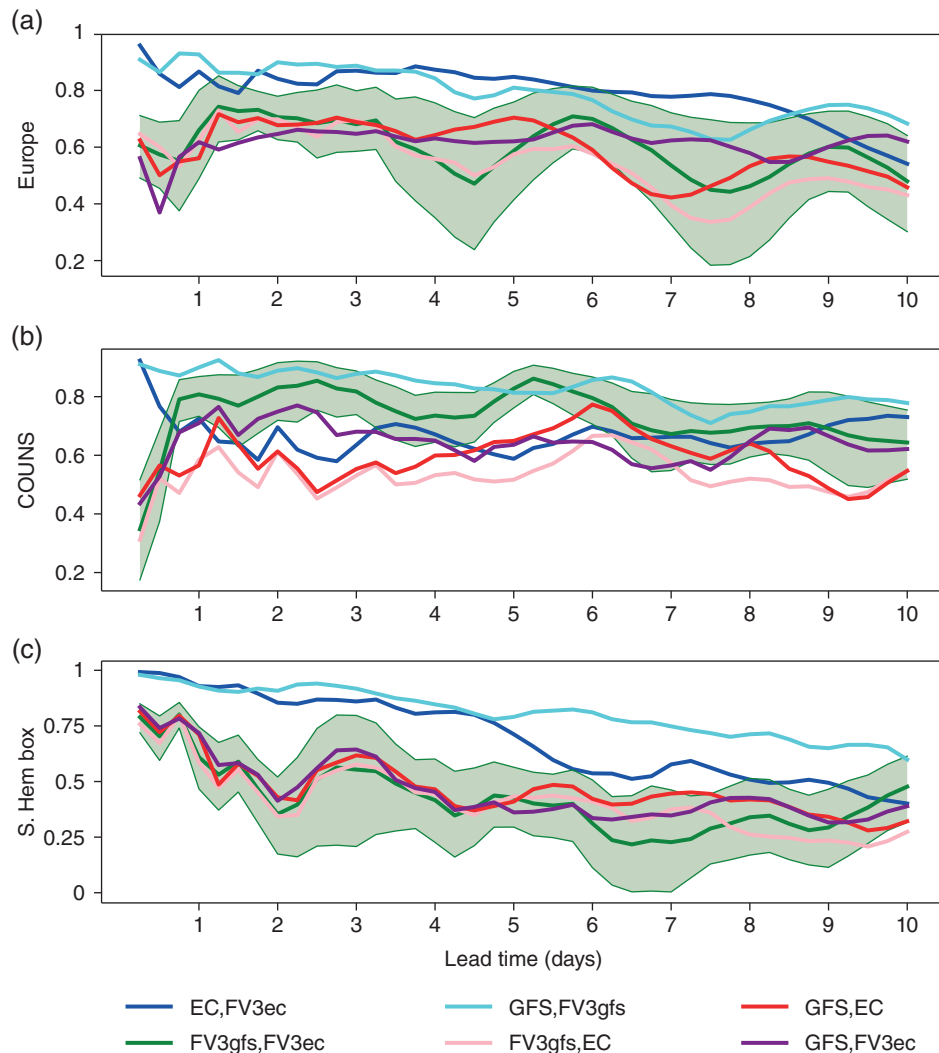
FIGURE 7 Correlation of errors between different experiments (see legend) versus lead time, including confidence intervals from 5th to 95th percentile based on a bootstrap method for FV3gfs versus FV3ec (green shading), for (a) Europe, (b) CONUS and (c) SE Pacific

We have also diagnosed orographic drag in the model to further understand the results.

The relative importance is addressed by comparing forecasts produced with the same initial conditions but different models and forecasts using different initial conditions but the same model. The models used for this study are the ECMWF-IFS, NCEP-GFS and GFDL-fvGFS with similar horizontal resolution (9 km for the first and 13 km for the others). While the first two models have a hydrostatic and spectral dynamical core, the fvGFS is non-hydrostatic with a cube-sphere grid and a finite-volume discretization. At the same time the physical parametrizations are similar for many of the processes in GFS and fvGFS. The GFDL-fvGFS has been initialized from both ECMWF and NCEP initial conditions. The caveat with this approach is that the model plays a significant role in the creation of initial conditions, and some of the difference in the initial conditions is therefore due to model differences.

The main conclusion is that the initial conditions play the major role for differences between the tested configurations for RMSE of z500 for at least 8 days into the forecasts, and for 5 days for precipitation. For the mean error (bias), the model formulation is the dominating factor for the investigated parameters. For the day-to-day variation of RMSE for z500, the initial conditions dominate for Europe and for the Southern Hemisphere, while for North America the model formulation has a larger influence.

For the RMSE the forecasts using ECMWF initial conditions yield the lowest RMSE in the medium range for both Northern and Southern Hemispheres, with similar scores for the ECMWF forecasts and the fvGFS model when using ECMWF initial conditions. We also find that the forecasts initialized from ECMWF analyses have an advantage for rainfall predictions. This result should not directly be interpreted to mean the scores are insensitive to the model formulation,
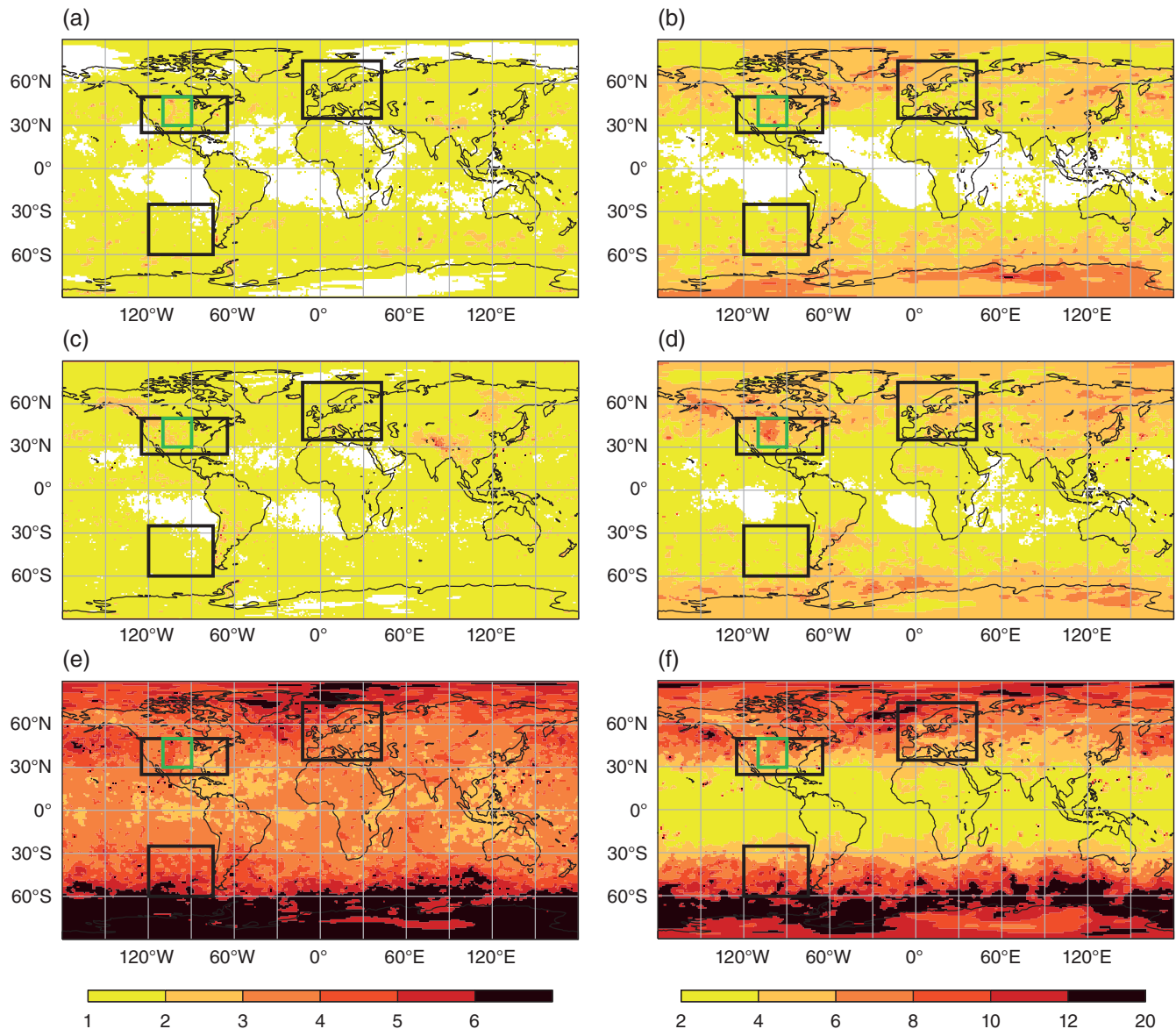
FIGURE 8 RMS difference between (a, b) FV3gfs and GFS, (c, d) FV3ec and EC and (e, f) FV3ec and FV3gfs at (a, c, e) 6 hr and (b, d, f) 30 hr lead time. The black boxes outline the areas used in Figure 7. The green sub-box over US is explained in the text

but rather that the two different models have similar quality although they are very different in formulation.

For the day-to-day variation of the errors, relatively large correlations are found for all pairs of experiments. This is not unexpected as the flow conditions affect all forecasts and during unpredictable situations all forecasts are more likely to have large errors. Such situations vary with season and were recently reviewed by Lillo and Parsons (2016). On top of the correlations given by the flow situation, the initial conditions contribute most to the error correlations over Europe and the Southern Hemisphere. The relatively large difference between the two used sets of initial conditions over oceans is one contributing factor. Such differences will amplify quickly, especially over regions with strong baroclinicity such as the Atlantic and southern oceans. The experimentation here also

confirmed that errors in the initial conditions played the major role for the March 2016 ECMWF forecast bust discussed in Magnusson (2017) and Grams *et al.* (2018).

For North America the picture is different regarding the dominating factor for the day-to-day variability. The strongest correlations are still found between the errors from GFS and FV3gfs, but less between EC and FV3ec. The stronger correlation for the first pair might also have been influenced by the similarities in the physical parametrization packages used in the GFDL-fvGFS and GFS models. The results show that difference between EC and FV3ec start to appear in the lee of the Rockies, and we find the fastest separation during passages of troughs over the mountain ridge. As we find large differences between the way the two models handle orographic drag, this is a plausible explanation for the forecast differences. The
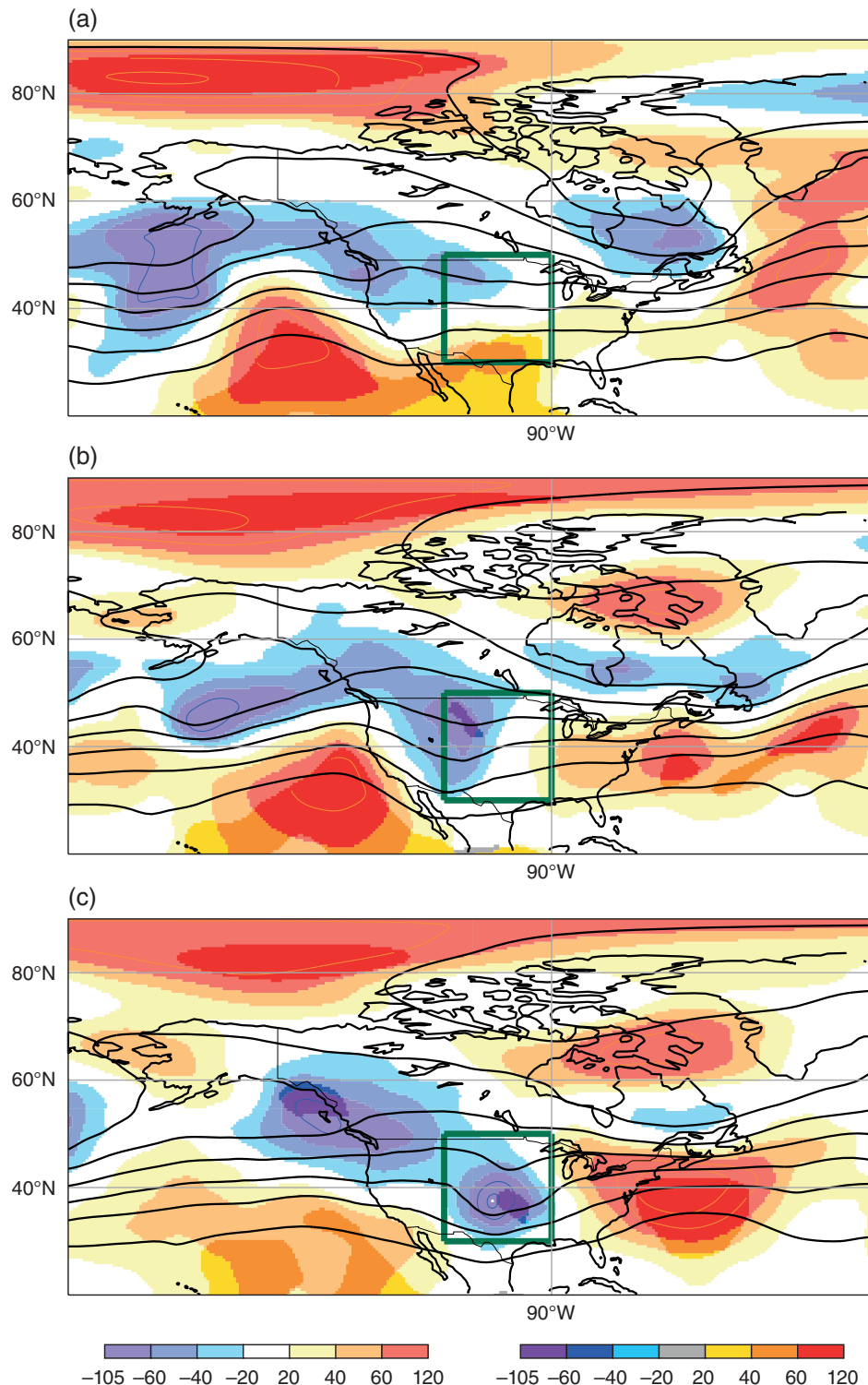
FIGURE 9 Composites of z500 (contours) and anomalies (shading) for the analyses for the top 10% of cases of dRMSD between Fv3ec and EC and between +6 hr and +30 hr. The composites are for (a) 6 hr before, (b) 18 hr after, and (c) 42 hr after the initialization times for the forecasts

difference in orographic drag between numerical weather prediction models is more widely discussed in Zadra (2015) and Sandu *et al.* (2016). This result suggests that special attention should be paid to perturbing the drag processes in ensemble forecasting systems to capture the flow-dependent forecast uncertainties. The current developments for model uncertainties at ECMWF are outlined in Leutbecher *et al.* (2017). From a forecast user perspective, being aware of situations where the differences in the model itself plays a larger role is useful when interpreting results from different forecasting systems.
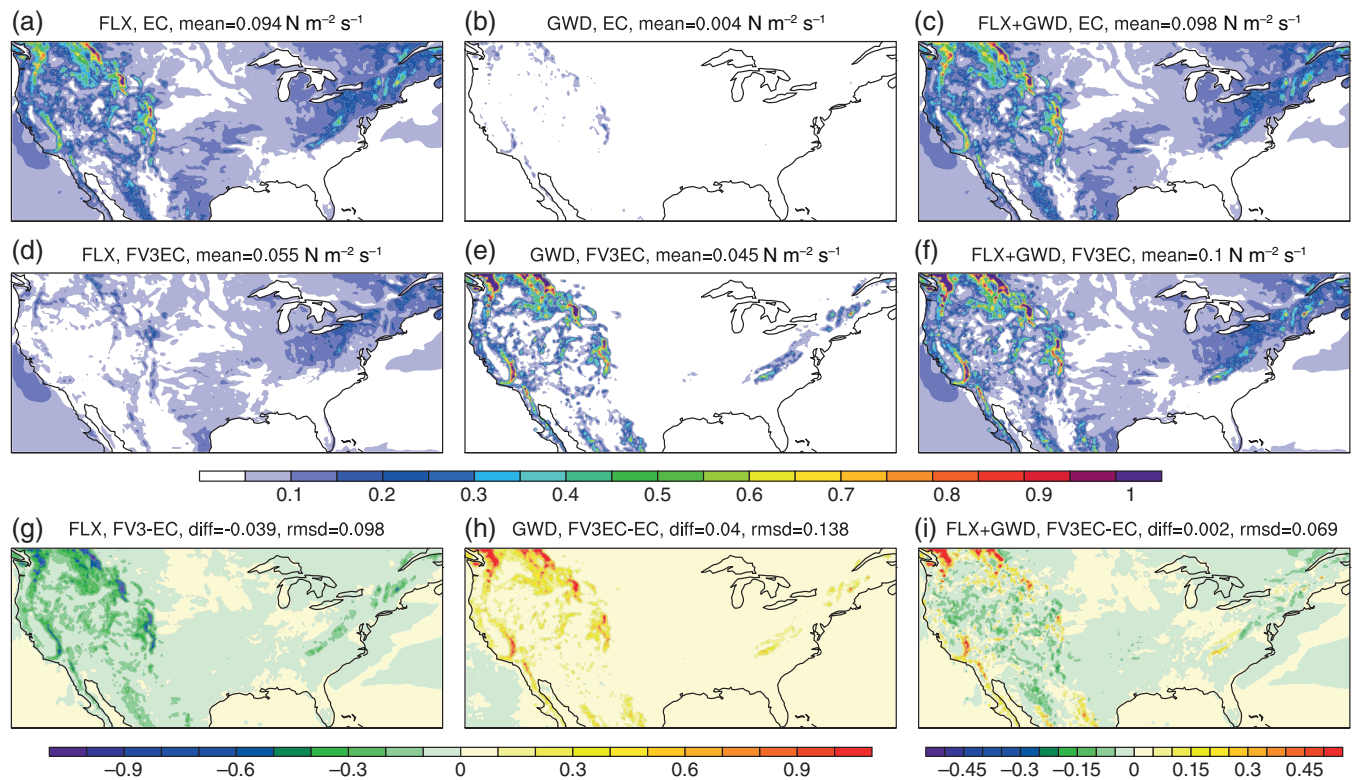
FIGURE 10 Surface stress averaged over days 0–5 in all forecasts from (a, d) turbulence and (b, e) orographic gravity waves and flow blocking and (c, f) their sum for (a)–(c) EC, (d)–(f) FV3ec and (g)–(i) their difference

Although the initial conditions seem to be most important for medium-range error differences between the experiments, we have to stress that the model plays a very important role in the creation of the initial conditions by the use of a short forecast as "first guess" in the data assimilation, as well as being used for the nonlinear outer trajectory and the tangent-linear model in the 4D-Var minimization. We also believe that the model formulation is much more important for the errors in weather parameters, such as 2 m temperature, cloud cover and 10 m wind speed. Further work using this framework will investigate the error for such parameters.

REFERENCES

Balsamo, G., Agusti-Panareda, A., Albergel, C., Beljaars, A., Boussetta, S., Dutra, E., Komori, T., Lang, S., Munoz-Sabater, J., Pappenberger, F., de Rosnay, P., Sandu, I., Wedi, N., Weisheimer, A., Wetterhall, F. and Zsótér, E. (2014) *Representing the Earth surfaces in the integrated forecasting system: Recent advances and future challenges.* Technical Report 729, ECMWF, Reading, UK.

Bauer, P., Thorpe, A.J. and Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525, 47–55. https://doi.org/10.1038/nature14956.

Bechtold, P., Kohler, M., Jung, T., Doblas-Reyes, F., Leutbecher, M., Rodwell, M.J., Vitart, F. and Balsamo, G. (2008) Advances in simulating atmospheric variability with the ECMWF model: from synoptic to decadal time-scales. *Quarterly Journal of the Royal Meteorological Society*, 134, 1337–1351.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.P., Beljaars, A. and Bormann, N. (2014) Representing equilibrium and non-equilibrium convection in large-scale models. *Journal of the Atmospheric Sciences*, 71(2), 734–753.

Beljaars, A., Brown, A.R. and Wood, N. (2004) A new parametrization of turbulent orographic form drag. *Quarterly Journal of the Royal Meteorological Society*, 130, 1327–1347.

Bonavita, M., Isaksen, L. and Hólm, E.V. (2012) On the use of EDA background-error variances in the ECMWF 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, 138, 1540–1559. https://doi.org/10.1002/qj.1899.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D., Ebert, E., Fuentes, M., Hamill, T., Mylne, K.R., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Silva Dias, P., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L. and Worley, S. (2010) THORPEX Interactive Grand Global Ensemble (TIGGE). *Bulletin of the American Meteorological Society*, 91, 1059–1072.

Buehner, M., Morneau, J. and Charette, C. (2013) Four-dimensional ensemble-variational data assimilation for global deterministic

weather prediction. *Nonlinear Processes in Geophysics*, 20(5), 669–682. https://doi.org/10.5194/npg-20-669-2013.

Chen, J.-H., Chen, X., Lin, S.-J., Magnusson, L., Bender, M., Zhou, L. and Rees, S. (2018a) Tropical cyclones in GFDL fvGFS – impacts of dycore, physics and initial conditions, In *Proceedings of 33rd Conference on Hurricane and Tropical Meteorology*, Ponte Vedra, FL. American Meteorological Society, Boston, MA. https://ams.confex.com/ams/33HURRICANE/webprogram/Paper 339827.html; accessed 18 April 2019.

Chen, J.H., Lin, S.J., Zhou, L., Chen, X., Rees, S., Bender, M. and Morin, M. (2018b) Evaluation of tropical cyclone forecasts in the next generation global prediction system (NGGPS). *Monthly Weather Review*.

Chen, S., Hong, Y., Gourley, J.J., Huffman, G.J., Tian, Y., Cao, Q., Yong, B., Kirstetter, P., Hu, J., Hardy, J., Li, Z., Khan, S.I. and Xue, X. (2013a) Evaluation of the successive v6 and v7 TRMM multisatellite precipitation analysis over the continental United States. *Water Resources Research*, 49(12), 8174–8186. https://doi.org/10.1002/2012WR012795.

Chen, X., Andronova, N., Van Leer, B., Penner, J.E., Boyd, J.P., Jablonowski, C. and Lin, S.J. (2013b) A control-volume model of the compressible Euler equations with a vertical Lagrangian coordinate. *Monthly Weather Review*, 141(7), 2526–2544. https://doi.org/10.1175/MWR-D-12-00129.1.

Chun, H.Y. and Baik, J.J. (1994) Weakly nonlinear response of a stably stratified atmosphere to diabatic forcing in a uniform flow. *Journal of the Atmospheric Sciences*, 51(21), 3109–3121. https://doi.org/10.1175/1520-0469(1994)051<3109:WNROAS>2.0.CO;2.

Clough, S.A., Shephard, M.W., Mlawer, E.J., Delamere, J.S., Iacono, M.J., Cady-Pereira, K., Boukabara, S. and Brown, P.D. (2005) Atmospheric radiative transfer modeling: a summary of the AER codes. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 91, 233–244. https://doi.org/10.1016/j.jqsrt.2004.05.058.

ECMWF. (2016) *IFS Documentation CY41R2*. Technical report, ECMWF, Reading, UK.

Ek, M.B., Mitchell, K.E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G. and Tarpley, J.D. (2003) Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. *Journal of Geophysical Research: Atmospheres*, 108(D22). https://doi.org/10.1029/2002JD003296.

Forbes, R.M. and Ahlgrimm, M. (2014) On the representation of high-latitude boundary-layer mixed-phase cloud in the ECMWF global model. *Monthly Weather Review*, 142(9), 3425–3445.

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. *Quarterly Journal of the Royal Meteorological Society*, 144, 2577–2591. https://doi.org/10.1002/qj.3353.

Han, J. and Pan, H.L. (2011) Revision of convection and vertical diffusion schemes in the NCEP global forecast system. *Weather and Forecasting*, 26(4), 520–533. https://doi.org/10.1175/WAF-D-10-05038.1.

Harrison, M.S.J., Palmer, T.N., Richardson, D.S. and Buizza, R. (1999) Analysis and model dependencies in medium-range ensembles: two transplant case-studies. *Quarterly Journal of the Royal Meteorological Society*, 125, 2487–2515. https://doi.org/10.1002/qj.49712555908.

Hazelton, A.T., Harris, L. and Lin, S.J. (2018) Evaluation of tropical cyclone structure forecasts in a high-resolution version of the multiscale GFDL fvGFS model. *Weather and Forecasting*, 33(2), 419–442. https://doi.org/10.1175/WAF-D-17-0140.1.

Holton, J.R. (1993) *An Introduction to Dynamic Meteorology*. Academic Press, Cambridge, MA.

Hong, S.Y., Noh, Y. and Dudhia, J. (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly Weather Review*, 134(9), 2318–2341. https://doi.org/10.1175/MWR3199.1.

Kim, Y.J. and Arakawa, A. (1995) Improvement of orographic gravity wave parameterization using a mesoscale gravity wave model. *Journal of the Atmospheric Sciences*, 52(11), 1875–1902. https://doi.org/10.1175/1520-0469(1995)052<1875:IOOGWP>2.0.CO;2.

Kim, Y.-J. and Doyle, J.D. (2005) Extension of an orographic-drag parametrization scheme to incorporate orographic anisotropy and flow blocking. *Quarterly Journal of the Royal Meteorological Society*, 131, 1893–1921. https://doi.org/10.1256/qj.04.160.

Leutbecher, M., Lock, S., Ollinaho, P., Lang, S.T.K., Balsamo, G., Bechtold, P., Bonavita, M., Christensen, H.M., Diamantakis, M., Dutra, E., English, S.J., Fisher, M., Forbes, R.M., Goddard, J., Haiden, T., Hogan, R.J., Juricke, S., Lawrence, H., MacLeod, D., Magnusson, L., Malardel, S., Massart, S., Sandu, I., Smolarkiewicz, P.K., Subramanian, A., Vitart, F., Wedi, N. and Weisheimer, A. (2017) Stochastic representations of model uncertainties at ECMWF: state of the art and future vision. *Quarterly Journal of the Royal Meteorological Society*, 143, 2315–2339. https://doi.org/10.1002/qj.3094.

Lillo, S.P. and Parsons, D.B. (2016) Investigating the dynamics of error growth in ECMWF medium-range forecast busts. *Quarterly Journal of the Royal Meteorological Society*, 143, 1211–1226. https://doi.org/10.1002/qj.2938.

Lin, S.J. (1997) A finite-volume integration method for computing pressure gradient force in general vertical coordinates. *Quarterly Journal of the Royal Meteorological Society*, 123, 1749–1762. https://doi.org/10.1002/qj.49712354214.

Lin, S.J. (2004) A 'vertically Lagrangian' finite-volume dynamical core for global models. *Monthly Weather Review*, 132(10), 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2.

Lorenz, E.N. (1982) Atmospheric predictability experiments with a large numerical model. *Tellus*, 34, 505–513.

Lott, F. and Miller, M.J. (1997) A new subgrid-scale orographic drag parametrization: its formulation and testing. *Quarterly Journal of the Royal Meteorological Society*, 123, 101–127.

Magnusson, L. (2017) Diagnostic methods for understanding the origin of forecast errors. *Quarterly Journal of the Royal Meteorological Society*, 143, 2129–2142. https://doi.org/10.1002/qj.3072.

Magnusson, L. and Källén, E. (2013) Factors influencing skill improvements in the ECMWF forecasting system. *Monthly Weather Review*, 141, 3142–3153.

Morcrette, J.-J., Barker, H., Cole, J., Iacono, M. and Pincus, R. (2008) Impact of a new radiation package, NCRAD, in the ECMWF integrated forecasting system. *Monthly Weather Review*, 136(12), 4773–4798.

Orr, A., Bechtold, P., Scinocca, J., Ern, M. and Janiskova, M. (2010) Improved middle atmosphere climate and forecasts in the ECMWF model through a non-orographic gravity wave drag parameterization. *Journal of Climate*, 23(22), 5905–5926.

Polland, R.T., Rhines, P.B. and Thompson, R. (1973) The deepening of the wind-mixed layer. *Geophysical Fluid Dynamics*, 3, 381–404.

Putman, W.M. and Lin, S.J. (2007) Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, 227, 55–78. https://doi.org/10.1016/j.jcp.2007.07.02.

Rabier, F., Järvinen, H., Klinker, E., Mahfouf, J.-F. and Simmons, A.J. (2000) The ECMWF operational implementation of four-dimensional variational assimilation. I: experimental results with simplified physics. *Quarterly Journal of the Royal Meteorological Society*, 126, 1143–1170. https://doi.org/10.1002/qj.49712656415.

Richardson, D.S. (2001) Ensembles using multiple models and analyses. *Quarterly Journal of the Royal Meteorological Society*, 127, 1847–1864. https://doi.org/10.1002/qj.49712757519.

Rodwell, M.J., Magnusson, L., Bauer, P., Bechtold, P., Cardinali, C., Diamantakis, M., Källén, E., Klocke, D., Lopez, P., McNally, A.P., Persson, A., Prates, F. and Wedi, N. (2012) Characteristics of occasional poor medium-range forecasts for Europe. *ECMWF Newsletter*, 131, 11–15. https://doi.org/10.21957/lnhbsxr4.

Rodwell, M.J., Magnusson, L., Bauer, P., Bechtold, P., Bonavita, M., Cardinali, C., Diamantakis, M., Earnshaw, P., Garcia-Mendez, A., Isaksen, L., Källén, E., Klocke, D., Lopez, P., McNally, A.P., Persson, A., Prates, F. and Wedi, N. (2013) Characteristics of occasional poor medium-range weather forecasts for Europe. *Bulletin of the American Meteorological Society*, 94, 1393–1405.

Sandu, I., Bechtold, P., Beljaars, A., Bozzo, A., Pithan, F., Shepherd, T.G. and Zadra, A. (2016) Impacts of parameterized orographic drag on the Northern Hemisphere winter circulation. *Journal of Advances in Modeling Earth Systems*, 8(1), 196–211. https://doi.org/10.1002/2015MS000564.

Sandu, I., Beljaars, A. and Balsamo, G. (2014) Improving the representation of stable boundary layers. *ECMWF Newsletter*, 138, 24–29.

Sela, J.G. (2009) *The implementation of the sigma pressure hybrid coordinate into the GFS*. Office Note 461, NCEP, Camp Springs, MD.

Sela, J.G. (2010) *The derivation of sigma pressure hybrid coordinate semi-Lagrangian model equations for the GFS*. Office Note 462, NCEP, Camp Springs, MD.

Tiedtke, M. (1993) Representation of clouds in large-scale models. *Monthly Weather Review*, 121(11), 3040–3061.

Torn, R.D. and Hakim, G.J. (2008) Ensemble-based sensitivity analysis. *Monthly Weather Review*, 136(2), 663–677. https://doi.org/10.1175/2007MWR2132.1.

Untch, A. and Hortal, M. (2004) A finite-element scheme for the vertical discretization of the semi-Lagrangian version of the ECMWF forecast model. *Quarterly Journal of the Royal Meteorological Society*, 130, 1505–1530. https://doi.org/10.1256/qj.03.173.

Zadra, A. (2015) *WGNE drag project: an inter-model comparison of surface stresses*. Technical report, Environment Canada, Dorval QC, Canada. http://collaboration.cmc.ec.gc.ca/science/rpn/drag_project/documents/wgne_drag_project_report01.pdf; accessed 18 April 2019

Zhao, Q. and Carr, F.H. (1997) A prognostic cloud scheme for operational NWP models. *Monthly Weather Review*, 125(8), 1931–1953. https://doi.org/10.1175/1520-493(1997)125<1931:APCSFO>2.0.CO;2.

Zhou, L., Lin, S.J., Chen, J.H., Harris, L.M., Chen, X. and Rees, S.L. (2019) Toward convective-scale prediction within the next generation global prediction system. *BAMS*. https://doi.org/10.1175/BAMS-D-17-0246.1.