Machine Learning for Model Error Inference and Correction

Massimo Bonavita[1] **and Patrick Laloyaux**[1]

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

Key Points:

- Model error is currently the biggest barrier to improve forecast accuracy in Weather and Climate Prediction
- An Artificial Neural Network (ANN) is trained to simulate the error in the ECMWF operational model
- Using the model error from the ANN inside weak-constraint 4D-Var extends the model error correction to the whole atmospheric column

Corresponding author: Massimo Bonavita, massimo.bonavita@ecmwf.int

Abstract

Model error is one of the main obstacles to improved accuracy and reliability in state-of-the-art analysis and forecasting applications, both in Numerical Weather Prediction (NWP) and in Climate Prediction, conducted with comprehensive high resolution General Circulation Models. In a data assimilation framework, recent advances in the context of weak constraint 4D-Var have shown that it is possible to estimate and correct for a large fraction of systematic model error which develops in the stratosphere over short-range forecast ranges. The recent explosion of interest in Machine Learning/Deep Learning technologies has been driven by their remarkable success in disparate application areas. This raises the question of whether model error estimation and correction in operational NWP and Climate Prediction can also benefit from these techniques. In this work, we aim to start to give an answer to this question. Specifically, we show that Artificial Neural Networks (ANN) can reproduce the main results obtained with weak constraint 4D-Var in the operational configuration of the IFS model of ECMWF. We show that the use of ANN models inside the weak-constraint 4D-Var framework has the potential to extend the applicability of the weak constraint methodology for model error correction to the whole atmospheric column. Finally, we discuss the potential and limitations of the Machine Learning/Deep Learning technologies in the core NWP tasks. In particular, we reconsider the fundamental constraints of a purely data driven approach to forecasting and provide a view on how to best integrate Machine Learning technologies within current data assimilation and forecasting methods.

Plain Language Summary

Model error is one of the main obstacles to improved accuracy and reliability in current Numerical Weather Prediction and in Climate Prediction. Recent advances in Data Assimilation at ECMWF indicate that it is possible to estimate and correct for a large fraction of systematic model error in the stratosphere. The question we address is whether Machine Learning techniques can be used alone and in conjunction with standard Data Assimilation methods to improve on those results. We show that it is indeed possible to extend current Data Assimilation capabilities in operational state-of-the-art forecast systems using Machine Learning tools and we discuss the potential and limitations of future applications of these ideas to other core NWP tasks.

1 Introduction

Numerical Weather Prediction (NWP) can be seen as an initial value problem where a numerical model is integrated in time to forecast the future state of the atmosphere and, increasingly, of the other components of the Earth System that interact with it. Like any other forecasting enterprise, NWP forecasts are affected by errors. In the data assimilation community, forecast errors are traditionally partitioned between errors from evolved erroneous initial conditions and model errors. This distinction is, for example, formalised in the evolution equation of the state error covariance in the Kalman Filter (Kalman, 1960)

$$\mathbf{P}_t^b = \mathbf{M}\mathbf{P}_{t-1}^a\mathbf{M}^{\mathrm{T}} + \mathbf{Q}_t, \tag{1}$$

where the state error covariance matrix of the background forecast state $\mathbf{P}_t^b$ is written as the sum of the evolved analysis errors from the previous analysis update ($\mathbf{M}\mathbf{P}_{t-1}^a\mathbf{M}^{\mathrm{T}}$ where $\mathbf{M}$ is the linear/linearised model) and a zero-mean stochastic model error of covariance $\mathbf{Q}_t$. This distinction has proved useful, as most data assimilation algorithms in current use can be seen as variations/extensions of the Kalman Filter, but it is also limited by significant assumptions: a) model error is assumed additive; b) model error is assumed to be white in time and c) model error is assumed to be zero-mean.

Assumptions a) and b) are somewhat relaxed in operational settings. For example, at ECMWF the model error parameterisations used in the Ensemble of Data Assimilations (EDA) to simulate model error evolution are based on a multiplicative ansatz (Buizza et al., 1999) and spatial model error correlations are cycled from one assimilation update to the next (Leutbecher et al., 2017). The third assumption (zero-mean errors) is probably the most important as it effectively makes any Kalman Filter based data assimilation system blind to the presence of systematic model errors (Dee, 2005). Note that we here use the term bias in a wider sense than it is typically used in the meteorological literature: Biases are systematic errors that can vary in space, time and prevalent meteorological conditions. Thus, we can encounter different model biases in different locations, at different times of day or year, in different meteorological conditions and they can be also influenced by systematic errors arising from the interaction with other components of the Earth System

In many data assimilation systems used in operational NWP, model bias is not accounted for explicitly. Rather, common strategies aim at reducing the impact of model biases on the performance of the assimilation system. Recognising that the impact of model biases on the assimilation algorithm mainly comes through the observation-minus-background (O-B) residuals, these strategies typically involve a combination of: a) debiasing the O-B residuals, for example through variational bias correction techniques (Auligné et al., 2007), and b) inflating the estimates of the background forecast errors sampled from an ensemble data assimilation system run in parallel to the main, higher resolution, analysis system (Bonavita et al., 2012; Whitaker & Hamill, 2012). Both techniques have proved effective in improving the performance of the data assimilation and forecast systems, but it is obvious that they are partial, sub-optimal solutions to the model bias problem. In fact, bias correction of the O-B residuals implicitly assumes that all the systematic components of these residuals are due to observation (and observation operator) biases. While this can be a reasonable working assumption for a large number of satellite radiances, the fact that we still see systematic O-B errors in largely unbiased observing systems (e.g., radiosondes, radio occultation observations from the Global Positioning System, a.k.a. GPS-RO) in operational data assimilation statistics shows that this is not the case in general. This effect is also visible in modern reanalyses (Hersbach et al., 2020) where long-term temperature trends in the stratospheric analysis show discontinuities connected to the introduction or withdrawal of specific observing systems. Inflating the background errors is a standard tool in ensemble data assimilation to deal with all components of the forecast error that are not properly sampled by the assimilation system (Houtekamer & Zhang, 2016). This technique has also proved effective in reducing the total analysis mean square error (Raanes et al., 2019), but it is clearly a blunt tool for dealing with model error. More importantly, any change to the Kalman Gain matrix in a bias-blind assimilation system will still result in a biased, sub-optimal analysis (Dee, 2005).

Weak Constraint 4D-Var (WC-4DVar) is an extension of 4D-Var which explicitly attempts to take model error into account in the solution of the 4D-Var assimilation problem (Tremolet, 2006). In the forcing formulation of WC-4DVar implemented at ECMWF, this is done by extending the 4D-Var control variable with a model error tendency term which is evaluated during the 4D-Var minimisation and used in the subsequent first-guess integration to de-bias the model trajectory:

$$\mathbf{x}_k = \mathbf{M}_{k,k-1}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}, \qquad k = 1, .., N \tag{2}$$

$$\begin{aligned} J_{WC}(\mathbf{x}_0, \boldsymbol{\eta}) = & \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) \\ & + \frac{1}{2}\sum_{k=0}^{N}\left(\left(H(\mathbf{x}_k) - \mathbf{y}_k\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(H(\mathbf{x}_k) - \mathbf{y}_k\right)\right) + \frac{1}{2}\left(\boldsymbol{\eta} - \boldsymbol{\eta}^b\right)^{T} \mathbf{Q}^{-1}\left(\boldsymbol{\eta} - \boldsymbol{\eta}^b\right), \end{aligned} \tag{3}$$

where $\boldsymbol{\eta}$ is the model error forcing (this is kept constant over the assimilation window, which is the main approximation of the IFS implementation of WC-4DVar), $\boldsymbol{\eta}^b$ is the prior estimate of the model error forcing and $\mathbf{Q}$ is the model error covariance matrix.

While this WC-4DVar formulation has been used at ECMWF since 2009, it is only very recently (IFS Cycle 47R1, scheduled to become operational from July 2020) that WC-4DVar has been shown to be effective at correcting stratospheric model biases (Laloyaux, Bonavita, Dahoui, et al., 2020). The key insight of this revised WC-4DVar implementation has been to impose scale separation between the error covariance matrices describing the spatial structures of background error $\mathbf{B}$ and of model biases $\mathbf{Q}$ (see Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, & Gürol, 2020, for a detailed explanation). The scale separation allows to successfully de-alias initial state and model error corrections during the 4D-Var minimisation, and is consistent with a view that model biases represent a type of errors that take place on larger spatial and longer temporal scales than background errors. It is also apparent from Equations (2) and (3) that WC-4DVar estimates a model error tendency term which is then applied as an additional forcing term in the prognostic equations of the model. Thus, it can be viewed as a data-driven algorithm to estimate (some of) the missing physical forcing in the model prognostic equations. In other words, WC-4DVar as described in Equations (2) and (3) is a type of on-line machine learning algorithm.

Machine learning (ML) methods, and more specifically the Deep Learning (DL) implementations of ML, have seen a remarkable resurgence over the past decade (Chollet, 2018). This was driven by the unrivalled results obtained through ML/DL technologies in a vast range of problems in computer vision, speech recognition, natural language processing and translation, among others (Goodfellow et al., 2016). At a fundamental level, most of the successful ML applications in use today implement a type of supervised statistical learning where we aim to learn from a dataset of examples $(\mathbf{X}, \mathbf{Y})$ a (possibly) nonlinear mapping between "features" $\mathbf{X} = \mathbf{x}^1, \ldots, \mathbf{x}^m$ and a corresponding set of observed targets ("labels") $\mathbf{Y}$. This is usually done by assuming a parametric model for the conditional distribution of the observations, $p_{\text{model}}(\mathbf{Y}|\mathbf{X}, \theta)$, and maximising the likelihood of the model over the empirical data distribution $p_{obs}(\mathbf{Y}|\mathbf{X})$

$$\boldsymbol{\theta}_{ML} = \arg\max_{\boldsymbol{\theta}} \; p_{\text{model}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}). \tag{4}$$

Under standard i.i.d. (independent and identically distributed) conditions for the features and observations distributions, Equation (4) can be transformed in the equivalent optimisation problem of maximising the log-likelihood of the predictive model under the observed distribution

$$\boldsymbol{\theta}_{ML} = \arg\max_{\theta} \; \sum_{i=1}^{m} \log \left(p_{\text{model}}(\mathbf{y}^i|\mathbf{x}^i, \boldsymbol{\theta}) \right), \tag{5}$$

where i is the index running over the m members of the examples' dataset. This is equivalent to minimising the cross-entropy between the two distribution (Goodfellow et al., 2016). For our purposes, we are interested in discovering a statistical regression law between model error (or, to be precise, available estimates of model error) and a set of predictors (features) to be defined based on physical intuition and experimental results. The simplest approach is assuming a linear relationship between $(\mathbf{Y}, \mathbf{X})$ represented by the affine transformation

$$\mathbf{Y} = \mathbf{W}\mathbf{X} + \mathbf{b}. \tag{6}$$

This is equivalent to assuming a Gaussian predictive model of the form

$$p_{\text{model}}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{Y}|\mathbf{W}\mathbf{X} + \mathbf{b}, \mathbf{I}), \tag{7}$$

where the general set of learnable parameters $\boldsymbol{\theta}$ has been particularised to the sets of weights $\mathbf{W}$ and bias coefficients $\mathbf{b}$ of a generic neural network. Maximising the log-likelihood (or, more commonly, minimising the negative log-likelihood) of this model

leads to the standard "Normal" equations. Adding constraints on the size of the regression coefficients matrix $\mathbf{W}$ (known in different communities as Tikhonov regularisation, ridge regression, weight decay) or the sparsity of said matrix (Tibshirani, 1996) can be seen as ways of improving the generalisation properties of the estimator by trading increased bias for reduced variance.

The main limitation of the regression model in Equation (6) lies in its limited capacity. If the underlying relation between $(\mathbf{Y}, \mathbf{X})$ is nonlinear, then the maximum likelihood estimator in Equation (5) will be sub-optimal. In our problem of model error estimation, it is a priori unclear how much of an issue this is. The WC-4DVar of Equation (3) is implemented at ECMWF in an incremental formulation, so it can deal with moderate nonlinearities through repeated re-linearisation steps (Bonavita et al., 2018). In Deep Learning, the nonlinearity problem is solved by introducing multiple additional layers in the regression that implement nonlinear transformations between their inputs and outputs (hidden layers). Even in their simplest algorithmic form, these nonlinear regressors variously known as Feedforward Neural Networks, Artificial Neural Networks (ANNs), or MultyLayer Perceptrons (MLPs) have the remarkable property of being universal function approximators (Cybenko, 1989). Thus, an ANN of sufficient capacity can theoretically learn any nonlinear mapping to any desired level of accuracy, given a sufficiently large and representative training dataset.

Attempts to use DL techniques to estimate and correct for model errors have already been documented in the geophysical literature. For example, Watson (2019) uses ANN to estimate model error tendencies in the Lorenz '96 system and uses them to correct short and long range forecasts with significant improvements both in forecast skill and model climate statistics. In that work an approximate (coarser) version of the Lorenz '96 model was still used for prediction, and the ANN was used to "fill in" the gaps with respect to the high resolution, "true" version of the model. This idea of hybridising machine learning methods with knowledge-based models is also exploited in the influential paper of Pathak et al. (2018), where a different ML technique is employed (Jaeger, 2001), but also very good results are obtained in two low-order models. In a similar vein, Bolton & Zanna (2019) present an oceanographic application of hybrid forecasting using Convolutional Neural Networks (CNN) in a simplified Ocean model. Again, the goal was to reproduce the effects of unresolved physical processes in a coarser version of their reference model. More recently, Brajard et al. (2020) demonstrate a way to combine ML with data assimilation of noisy and partial observations. In their scheme, DA and ML alternate in producing progressively more accurate estimates of the state and of the surrogate predictive model. This idea has been framed into a unifying Bayesian formalism by Bocquet et al. (2020), which allows to develop approximations and alternative algorithms.

In the works described above and, to the Authors' knowledge, in other recent relevant literature in the geophysical domain, the application of ML techniques for model error inference and correction has been studied in the context of low-order, simplified models. Thus, while the reported results appear encouraging, questions remain about the extent to which those results are applicable and relevant for high resolution, operational level data assimilation and forecasting applications. These applications pose a new set of additional challenges. Firstly, in real world applications the true state is typically unknown. What is known are incomplete and noisy estimates of the true state, either directly through observations (which are affected by random and systematic errors of their own) or indirectly through analyses produced by a data assimilation system (which are themselves affected by model and observation errors). Secondly, and possibly more importantly, the dimensions of the analysis and forecast system in operational NWP are very large. In the current Integrated Forecasting System (IFS) used at ECMWF, the size of the model state vector is $\mathcal{O}(10^{10})$ and the size of the analysis control vector is $\mathcal{O}(10^{8})$. These numbers are orders of magnitude

larger than those for typical low or intermediate complexity models discussed in the literature and they pose a new set of practical and conceptual questions. The aim of this work is to give some initial and, at this stage, necessarily tentative answers to these questions. The main conclusion that we derive from the results presented in the following is that, while a considerable amount of work still needs to be done, there is a concrete prospect to successfully integrate ML solutions inside the 4D-Var machinery of state-of-the-art operational NWP systems like the IFS and, by doing so, of significantly improving their analysis accuracy and their forecast skill.

This paper is organised as follows. In Section 2 we describe the ML methodology used in this work and the results achieved in terms of the predictive properties of the ML model. In Section 3 we examine the structure of the model error tendency predictions of the ML model and compare them to the predictions of the forthcoming operational version of WC-4DVar. In Section 4 we examine the results of using the ML-derived model error tendency predictions in cycled 4D-Var experiments, both as a stand-alone replacement of the WC-4DVar estimates and in conjunction with WC-4DVar. In Section 5 we discuss these results further in terms of their implications for our future research and, more generally, in the context of the current research effort to integrate ML tools in the NWP chain. Conclusions are offered in Section 6.

2 Machine Learning Methodology and Results

2.1 Set-up of the regression problem

The first task in a regression setting is to identify the set of predictors and predictands that are most relevant (in ML terminology, the examples $(\mathbf{X}, \mathbf{Y})$ of the supervised learning problem). As remarked in the introduction, in a real-world setting we do not have access to the true model error predictands ($\mathbf{Y}$), thus we need to find suitable substitutes. Generally speaking, the fundamental sources of information about model error are observations. In a data assimilation context, we can access this information directly in observation space (through background, O-B, and analysis, O-A, departures) or mediated by an analysis (through analysis increments fields, A-B). In this work we have chosen the second option, mainly because it is technically easier to implement, the increments have global, homogeneous coverage and are already available in the space of the IFS model variables: temperature (t), logarithm of surface pressure (lnsp), vorticity (vo), divergence (d), specific humidity (q)). We still think, however, that a direct use of observation departures would be a direction worth pursuing in the future. We remark here that this idea of using timeseries of analysis increments' fields to estimate the predictable component of model error is not new in the meteorological literature. For example, one of the algorithms proposed in Dee (2005) for the correction of model bias in a cycled data assimilation framework explicitly involves using an online model error estimate based on a running mean over past analysis increments (e.g., Eqs (43, 44) in Dee, 2005).

We can broadly consider two classes of predictors ($\mathbf{X}$). The first, which we call "climatological", comprises predictors that do not depend on the state of the flow. In this work, our climatological predictors are the set: (latitude, longitude, time_of_the_day, month). This set of predictors aims at capturing that part of model error which is related to geographical location, to the diurnal cycle and to the seasonal cycle. The other class of predictors used in this work are called "state" predictors. These are predictors that are meant to represent the part of model error linked to the large scale state of the flow, e.g. oceanic stratocumulus areas, Intertropical Convergence Zone, extra-tropical cyclonic areas, etc. In this first implementation, and with an operational application in mind, we have chosen the vertical columns of the background forecast fields of the subset of state variables of the model whose analysis increments are also used as predictands (i.e., t, lnsp, vo, d, q). This choice is practical,

but it can be potentially extended to other state variables and also to the use of state variables valid at different times. An example of possible avenues for expanding the set of state predictors is discussed in Section 2.2.

Connected with the choice of analysis and forecast fields as (a component of the) predictors and predictands, is the choice of horizontal spatial resolution for the fields whose vertical columns are used in the regression (see Figure 1 for a schematic of the ANN structure). In this work we have selected a resolution in spectral space of Triangular spectral truncation 21 (T21), which corresponds to an approximate grid spacing of 900 km on a reduced quadratic Gaussian grid. This choice is motivated by both practical and fundamental reasons. On the practical side, the coarse resolution chosen here facilitates the training phase of the ANN as it keeps its memory and computational requirements at a manageable level (in this work we did not have access to supercomputing resources for the training of the ANN). On the science side, this choice is motivated by the findings in Laloyaux, Bonavita, Dahoui, et al. (2020); Laloyaux, Bonavita, Chrust, & Gürol (2020) that only large scale model errors are predictable in a weak constraint 4D-Var framework. Additionally, there are fundamental arguments from ergodic dynamical systems theory that suggest that only large scale features of model error can be learned statistically. We will come back to these arguments in the discussion in Section 5.

2.2 Training the ANN

The training of the artificial neural networks (ANN) was conducted on a dedicated dual GPU workstation (NVIDIA Quadro GV100) using the open source deep learning backend Tensorflow (version 1.14.0, Abadi et al., 2016) and its high-level Python interface Keras. Initial experiments were conducted on an Intel i5 CPU-based workstation. The relative speed-up in the training phase achieved on the GPU system was a factor of approx. 3. The training dataset consisted of operational analysis increments and background forecasts collected over the whole year of 2018 every 36h (i.e., using one every three of the available analyses of the ECMWF operational 12-hourly assimilation cycle) at T21 resolution. The climatological predictors defined in Section 2.1 were extracted from the grib headers of the state predictors fields. The validation dataset was composed in the same way using a short two-months period from 1 January 2019. This dataset was used to get an indication of appropriate hyper-parameters' values. The test dataset used for verifying the performance of the ANN was composed by the analysis increments and background forecasts of a three and half month period starting on the 1 April 2019.

The statistical regressions have been computed separately for the three set of predictands and the related state predictors: mass (t, lnsp), wind (vo, d) and humidity (q), leading to three separate ANN models. The reason was again to reduce the computational and memory cost of the learning phase. This will be reviewed in the future, but we do not expect to see large benefits from performing a combined regression on the whole set of predictands as the statistical signatures of mass-wind error cross-correlations are typically small (Hamrud et al., 2015). We have tested two types of regression models. One is a standard linear regression with full connections between predictors and predictands. This implements the regression model in Equation 6. The number of trainable parameters is the number of input predictors times the number of output predictands (dimension of weight matrix $\mathbf{W}$) plus the number of predictands (bias vector $\mathbf{b}$). Considering for example the case of the full neural network for the mass variables (t, lnsp) with the current IFS number of vertical levels (137) this implies a number of trainable parameters equal to $142 \times 138 + 138 = 19734$. The number of vertical profiles in the training dataset is $\mathcal{O}(10^6)$, which, as we will see, is enough to prevent over-fitting. The other regression model is a nonlinear model where a nonlinear transformation is applied element-wise to the output of Equation (6) on

ANNs of increasing depth. The nonlinear transformation is modelled by the REctified Linear Unit (RELU) function, expressed by the function:

$$\mathrm{Relu}(\mathbf{x}) = \max(0, \mathbf{x}) \tag{8}$$

The nonlinear transformation in Equation (8) is applied to all layers of the ANN except the output layer. In the terminology we adopt in the following, relu_one_layer is the fully-connected ANN composed of two layers: an input layer where we use the nonlinear transform Equation 8 and a linear output layer. Similarly, relu_two(three)_layers refer to the fully connected ANN derived from relu_one_layer ANN through the addition of one (two) hidden nonlinear layer between input and output. In our terminology, the cardinality refers thus to the number of nonlinear layers in the ANN and not to the number of hidden layers of the model, as it is more common in the ML literature.

The minimiser used in the training is Adam (Kingma & Ba, 2014), which is an adaptive version of stochastic gradient descent (SGD). We found it to be generally able to show more monotonous convergence properties and require less tuning of its hyper-parameters (learning rate and decay rates) than standard SGD and other adaptive methods available in the Tensorflow toolbox. Regularisation is also an important aspect of deep learning methodology. In our case we found regularisation to be only moderately helpful, mainly improving monotonicity of convergence and slightly improving generalisation power of the model. After some trials, we have settled on weight decay for the linear regression model and dropout (Srivastava et al., 2014) with a dropout rate of 20% for the nonlinear models. This relative lack of sensitivity to regularisation methods is likely due to the relative shallowness of the ANN we have used and the fact that the size of our training dataset is one to two orders of magnitude larger than the number of model parameters.

As it is standard in regression settings, the Mean Square Difference between predicted and actual model errors is minimised during training. In order to give a more expressive view of the predictive capability of the ANN we present training results in terms of the coefficient of determination, which represents the proportion of total variance in the training sample that is explained by the model

$$R2 = 1 - \frac{SS_{red}}{SS_{tot}} \tag{9}$$

where $SS_{tot} = \sum_{i=1}^{m} (\mathbf{y}^i - \bar{\mathbf{y}})^2$ is the total Sum of Squares (proportional to the sample variance) and $SS_{res} = \sum_{i=1}^{m} (\mathbf{y}^i - \mathbf{f}^i)^2$ is the residual Sum of Squares. In a perfect model scenario where the model is able to accurately predict every instance of the sampling dataset, $R2 = 1$. As our model error generating processes are inherently stochastic, even a perfect model, i.e. a model that makes predictions sampling from the true error generating distribution, will produce some error, so that R2 will in general be smaller than 1 (this irreducible error is sometimes called Bayes error (Goodfellow et al., 2016)). Note also that a baseline model that always predicts the average value of the sampled predictand $\bar{\mathbf{y}}$ has a $R2 = 0$ and models that do worse than this baseline will have negative R2. The R2 coefficient has also been used in this work as stopping criterion in the training to avoid overfitting (i.e., training is stopped when R2 has not increased over the previous 20 epochs). In Figures 2, we present the results of the ANN training for the three sets of model error tendency predictands: (t, lnsp), (vo, d) and q. State and climatological predictors are used which means that the column background forecast fields as well as the metadata (latitude, longitude, time of the day, month of year) are selected as the input of the neural network. From this set of training and test results we draw the following conclusions:

- The mass errors (t, lnsp) are the most predictable: approximatelly 14% of the variance of the analysis increments of the test dataset is predicted by the best ANN model;

- The wind errors (vo, d) have lower predictability, with the best ANN accounting for $\sim 5\%$ of the variance of the analysis increments of the test dataset;
- The humidity errors (q) have the lowest predictability. Even the best ANN has a R2 not significantly larger than zero. This implies that it has no better predictive skill than a baseline model using the mean analysis increment of the training dataset;
- The predictive power of the ANNs increases going from linear to nonlinear models of increasing depth. The improvements are very large ($\sim 100\%$) going from the linear to the nonlinear regression with one nonlinear layer and saturate with the relu_three_layers model. Adding more nonlinear layers does not produce further improvements in test dataset R2 (not shown).

These results confirm that estimating model error in the IFS at the rather coarse scales we are considering here is a mildly nonlinear problem, which can partly explain the success of WC-4DVar in its current configuration (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, & Gürol, 2020). In the current WC-4DVar configuration only mass and (to a lesser extent) wind model errors are estimated and corrected, which also seems a good choice based on the results in Figure 2.

An interesting aspect of any regression model is to understand which of the predictors have the most predictive power. We have not looked into this aspect in great detail, but we have trained two separate regression models, one using only climatological predictors, the other only using state predictors. In Figure 3, we present results for the (t, lnsp) predictands. From this plot we can conclude that the state predictors are more informative than climatological predictors ($R2_{\text{state}} \sim 10\%$, $R2_{\text{climat}} \sim 8\%$) but both set of predictors contribute independent information to the final regression model ($R2_{\text{full}} \sim 14\%$).

2.3 Training the ANN with an Augmented State Predictor Set

There are several possible avenues for extending the set of predictors in our regression problem. One way would be to use the whole set of state variables considered in this work (t, lnsp, vo, d, q) as predictors in each regression problem. This would amount to try to leverage the cross-variable correlations in the model error estimates. In practice, mass-wind error cross-correlations are found to be small on average ($\sim 10\%$, Hamrud et al., 2015), so a considerably larger training dataset would likely be required to estimate these small covariances. In a similar vein, the set of state predictors could be augmented to include any other state variables that could potentially co-variate with the predictands, for example vertical velocity, precipitation rate, liquid water content, etc. Another option which we have started to investigate, is to extend the set of state predictors in time. The intuition here is to try to extract information on current errors not only from the forecast state valid at the same time but on its recent evolution. A simple and relatively inexpensive way of achieving this result is to augment the set of state predictors, which are 12-hour background forecast fields, with the analysis fields from which they were forecasted. This implies an approximate doubling of the size of the predictors (e.g., for (t, lnsp) from 142 to 280). An example of results from the training of ANNs using this flavour of augmented predictor set is shown in Figure 4. This plot suggests that the ANN trained on the augmented predictor set has more predictive power than the ANN trained on the standard set ($R2 \sim 15\%$ vs 14%). A similar improvement is seen for the wind error predictands ($R2 \sim 6\%$ vs 5%), while no improvement is visible in the humidity ANN results (not shown).

3 Predicting Model Error with Artificial Neural Networks

In this section we present a series of diagnostic results in order to give a first impression of what the model error tendencies predicted by the trained ANNs look like and how they compare with those estimated by WC-4DVar and also visible in observation departures. The plots presented in the following refer to one week of data but are indicative of the ANN results over the test period. Results shown here refer to relu_three_layers ANNs trained with the standard set of climatological and state predictors, not with the extended set described in Section 2.2. For comparison, we show weak-constraint 4D-Var model error estimates for an experiment run over the same period and initialised from the operational IFS.

In Figures 5a and 5b, we present a weekly average of the temperature model error tendencies estimated by the ANN (left) and by WC-4DVar (right). To be consistent with current IFS WC-4DVar practice, the ANN model error tendencies are derived from the ANN predicted analysis increments divided by the length (in hours) of the assimilation window. The current version of WC-4DVar is not active below model level 60 (approx. 100 hPa); the ANN is active everywhere and in the troposphere (below 100 hPa) it shows patterns of warm and cold error layers with larger intensities in the boundary layer. In the layer between model level 60 to 30 (approx. 100 to 10 hPa) both WC-4DVar and the ANN show a general tendency to warm the atmosphere, more noticeably in the tropics. This is consistent with the cold model bias seen in radiosonde temperature measurements in the lower troposphere (see below). Above model level 30 both ANN and WC-4Dvar show a generally negative (cooling) tendency, which is also consistent with the warm model bias with respect to radiosonde measurements in this layer of the model atmosphere. In the vorticity and divergence model error plots (Figures 5c-f) the corrections estimated by the ANN and WC-4DVar are smaller and more homogeneous. It is difficult to see clear physically interpretable patterns apart from a general tendency to decrease both parameters and the hint of a coherent negative-positive-negative divergence pattern in the tropical troposphere (Figure 5e). This last pattern appears to be a robust feature of the ANN regression (it is present in all the weekly averages computed over the test period, not shown), and thus it likely points to local issues in the current parameterised convection scheme, the data assimilation system, or both.

To obtain further insight in the spatial variability of the model error tendencies predicted by the ANN, we present in Figure 6 the weekly averaged plots of temperature model error tendencies from the ANN and WC-4DVar at model level 24 (approx. 5 hPa) and 50 (approx. 50 hPa). While the globally averaged values agree, the spatial structures are different: in particular, the ANN tendencies are larger scale and less intense than those of WC-4DVar. This is to be expected, as the WC-4DVar estimates are online estimates, and thus more sensitive to existing flow conditions than the ANN estimates.

It is also interesting to see the geographical distribution of the ANN-derived model error tendencies for model levels where the current WC-4DVar does not produce an estimate (i.e., below 100 hPa). An example is given in Figure 7 where the ANN estimates are shown for model levels close to 100 (a), 500 (b) and 850 hPa (c). From these plots one can see clear signatures of errors connected to downstream flow from the main mountain ranges (Rockies (b), Himalayas (a)), to convectively active areas (Amazons (b), Maritime Continent (b)), to storm tracks regions (south hemisphere storm tracks (b)) and to marine stratocumulus areas off the western seaboards of continental land masses (c). These features all point to predictable, flow-dependent errors in the model which the ANN regression tries to correct and they can be viewed as an additional model diagnostic tool.

Surface pressure is another component of the state vector whose model errors are not estimated by the current version of WC-4DVar but is an output of the ANN regression. Two examples of the ANN estimates of surface pressure errors are presented in Figure 8, one from a weekly average in July 2019 (a), the other from a week in November 2019 (b). It is interesting that while some features of the estimated surface pressure error appear stationary (e.g., oceanic western boundaries, convective areas like the maritime continent and the Amazons, etc.), seasonal variability is visible in other parts of the Globe where the underlying meteorology is significantly different and more sensitive to the seasonal cycle (e.g., Antarctic region, Siberian landmass). Again, each of these signals can provide potentially valuable diagnostic indications as they point to systematic problems in the short-range forecast and/or the use of observations in those areas. How these diagnostics produced by the applications of ANN and WC-4DVar compare with those derived from more traditional approaches based on the accumulation of data assimilation statistics (Rodwell & Jung, 2008) is an interesting line of research that we defer to future work.

4 Testing ANN in the IFS 4D-Var

While the analysis of model error predictions produced by an ANN can provide useful diagnostics of model error patterns and hint at their underlying drivers, a more stringent test of the ANN potential is whether its model error predictions can be gainfully used in a data assimilation context. We recall from Equations (2) and (3) that the current ECMWF WC-4DVar works by estimating a constant in time model error correction during the 4D-Var minimisation and then using that correction as a model forcing during the successive first guess trajectory integration. As explained in Section 2, the ANN used in this study has been trained to predict the analysis increments (A-B) of the operational IFS, whose systematic component can be viewed as an estimate of the cumulated model error over the 12 hour integration from one analysis update to the next. By dividing this quantity by the number of timesteps used in the 12 hour integration we obtain an estimate of the model error tendencies, under the same assumptions of time constancy over the assimilation window length of the ECMWF operational WC-4DVar. There are at least two main ways in which one can use the offline model error tendencies produced by the ANN. The first option, called NN_SC in the following, is based on a strong-constraint 4D-Var where the model first guess trajectories are corrected by the ANN model error tendencies (see Algorithm 1). The second option, called NN_WC, is based on the weak-constraint 4D-Var presented in the Equations (2) and (3) where the model forcing $\boldsymbol{\eta}$ is initialised by the ANN tendency (see Algorithm 2). The model forcing that comes out of the minimisation of Equation (3) which contains the ANN tendency updated by the weak-constraint 4D-Var minimisation is not carried forward in time. In the successive assimilation window, the model forcing $\boldsymbol{\eta}$ is initialised with the corresponding ANN tendency valid 12 hours later. It is also important to note that since current WC-4DVar is active only above 100 hPa, the model error tendencies below 100 hPa derive entirely from the ANN estimates in both NN_SC and NN_WC experiments. The results shown in the following come from assimilation and forecast experiments conducted with the latest available IFS cycle at the time of writing (Cycle 47R1, May 2020) and at the operational configuration for both the 4D-Var analysis step and the forecast step (TCo1279 spectral truncation, approx. 9 km grid spacing), over the period 16-07-2019 to 24-08-2019. The two experiments making use of the ANN model error estimates (NN_SC and NN_WC) are compared with an experiment using the standard operational weak constraint configuration (denoted "WC") and a strong-constraint 4D-Var used as a baseline.

Algorithm 1 NN_SC

Loop over the assimilation cycles from 16-07-2019 to 24-08-2019 :
Loop over the physical variables (T,lnsp), (vo,d) and q :
Compute the model error forcing with the trained ANN at valid times
Concatenate the outputs in a vector η^b
Minimise the strong constrain 4D-Var

$$J_{SC}(\mathbf{x}_0) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \frac{1}{2}\sum_{k=0}^{N}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right)$$

where the model trajectories are computed as

$$\mathbf{x}_k = \mathcal{M}_{k,k-1}(\mathbf{x}_{k-1}) + \boldsymbol{\eta}^b \qquad \text{for} \qquad k = 1, \ldots, N$$

Algorithm 2 NN_WC

Loop over the assimilation cycles from 16-07-2019 to 24-08-2019 :
Loop over the physical variables (T,lnsp), (vo,d) and q :
Compute the model error forcing with the trained ANN at valid times
Concatenate the outputs in a vector η^b
Minimise the weak-constrained 4D-Var

$$J_{WC}(\mathbf{x}_0, \boldsymbol{\eta}) = \frac{1}{2}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right)^{\mathrm{T}} \mathbf{B}^{-1}\left(\mathbf{x}_0 - \mathbf{x}_0^b\right) + \frac{1}{2}\sum_{k=0}^{N}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right)^{\mathrm{T}} \mathbf{R}_k^{-1}\left(\mathcal{H}_k(\mathbf{x}_k) - \mathbf{y}_k\right) + \frac{1}{2}\left(\boldsymbol{\eta} - \boldsymbol{\eta}^b\right)^{\mathrm{T}} \mathbf{Q}^{-1}\left(\boldsymbol{\eta} - \boldsymbol{\eta}^b\right)$$

where the model trajectories are computed as

$$\mathbf{x}_k = \mathcal{M}_{k,k-1}(\mathbf{x}_{k-1}) + \boldsymbol{\eta} \qquad \text{for} \qquad k = 1, \ldots, N$$

4.1 Time evolution of the model error estimates

Figure 9a shows the timeseries of the global mean model error correction estimated with weak-constraint 4D-Var between 15 and 24 August 2019. The model error is initialised at the beginning of the experiment from operations and is then cycled over the 12-hour assimilation windows. Weak constraint 4D-Var corrects the warm bias in the upper stratosphere and the cold bias in the mid/lower stratosphere. It correctly captures the transition layer (20 to 10 hPa) where the model bias changes from cold to warm. Figures 9b and 9c present the same diagnostic for NN_SC and NN_WC respectively. The transition level between the cold and the warm bias layers is estimated at the same pressure level as in weak-constraint 4D-Var. The main difference is in the upper stratosphere where the neural network produces a positive correction around 2hPa as this signal is present in the analysis increments used to train the neural network.

4.2 Evaluation of mean errors

One of the main successes of the new WC-4DVar introduced in IFS Cycle 47R1 has been the drastic reduction (up to 50%) of temperature biases in the ECMWF stratospheric analyses (Laloyaux, Bonavita, Dahoui, et al., 2020). The first set of diagnostics presented in Figure 10 aims at understanding if and in what measure these results can be achieved using an ANN (alone or as a first guess). The general impression from these plots is that the two ANN-driven WC-4DVar experiments produce similar results to one another and manage to broadly replicate the effects of current IFS WC-4DVar (though a closer look points to a relative better behaviour of WC-4DVar at certain heights/pressure levels, e.g. radiosonde temperature at 5 hPa). Apart from temperatures, current WC-4DVar also corrects for wind model errors in the stratosphere. The mean wind observation departures presented in Figure 10c confirm that WC-4DVar is able to correct for systematic wind model errors above 50 hPa, where it is fully active, while results are mixed in the transition layer between 100 and 50 hPa. The two ANN driven experiments show similar results in the layer above 50 hPa, but are also able to substantially reduce model biases in the atmospheric column down to approx. 700 hPa. In Figure 10e we present results for surface observations. The only significant differences are seen here in the departures for surface pressure observations. For these pressure observations, background departures for strong and weak constraint 4D-Var are almost identical, which is not surprising since the mass adjustments that current WC-4DVar performs in the stratosphere have small impact on the total column weight. On the other hand, the ANN driven WC-4DVar experiments show an approximate halving of the surface pressure observed biases, which confirms that the ANN derived corrections are also effective in the troposphere and for the surface pressure field.

4.3 Evaluation of random errors

The ability of WC-4DVar and its ANN-driven variants to effectively debias the first guess trajectories should in principle improve the successive analyses and forecasts by allowing the assimilation to make better use of the available observations. To investigate this aspect of the assimilation system performance we start by showing in Figure 11 the normalised standard deviation of analysis (O-A) and background (O-B) departures for AMSU-A radiances (a) and radiosonde temperature observations (b). The AMSU-A radiometer on board multiple operational and research meteorological satellites is a microwave radiometer whose channels are sensitive to deep layers of the atmosphere. The channels used in the experiments have weighting functions that peak in the troposphere (ch. 5 to 8) and the stratosphere (ch. 9 to 14). From the AMSU-A plot it is apparent that current WC-4DVar is more effective in the stratosphere and upper troposphere than either of the ANN-driven 4DVars, and equivalent

to the NN_WC set-up in the middle to lower troposphere (the NN_SC setup performs consistently worse). The picture is more nuanced for radiosonde observations, where the ANN 4DVars appear to perform significantly better than current weak and strong constraint 4DVar in the lower troposphere and comparably above. The results from other independent observing systems sensitive to atmospheric temperature not shown here (e.g., hyperspectral sounders) appear to confirm the advantage of the current WC-4DVar in the middle and lower stratosphere and suggest an improved performance of NN_WC in the middle and lower troposphere. We note, additionally, that the results from the stratospheric-peaking satellite radiances need to be taken with some caution, as they are influenced by the evolution of the corresponding bias correction coefficients. For the experiments reported in this work the bias correction coefficients have been initialised by a long-running pre-operational WC-4DVar, and thus are likely to be sub-optimal for the other configurations over most or all of the test period. In fact, no degradations are apparent for the ANN driven experiments when the verification is conducted against non-bias corrected observing systems (radiosondes, GPS-RO). For conventional wind observations (Figure 11c) there is hardly any significant difference among the four experiments. For satellite atmospheric motion vector winds (Figure 11d) the differences are clearer: the two ANN driven experiments improve results over standard weak and strong constraint 4D-Var in the boundary layer and in the Upper Troposphere Lower Stratosphere (UTLS) layer. This is consistent with results seen in the mean wind errors plots (Figure 10c 10d). The statistics for random errors affecting humidity sensitive observations are not presented as they generally do not show appreciable differences among the experiments. This is to be expected because neither the current WC-4DVar nor the two ANN-driven versions apply any additional model forcing for humidity, thus any change in behaviour would only be an indirect effect of changes to temperature/wind evolution.

On the other hand, there are significant changes in the diagnosed background error standard deviations for surface pressure observations (Figure 11e). Consistently with the results for the mean errors, the NN_WC version of ANN-driven 4D-Var significantly reduces random errors for surface pressure observations with respect to the reference strong and weak constraint 4D-Var. It is also to be noticed that NN_WC uses 6% more Dribu (Drifting Buoys) observations than the reference SC, due to the fact that more observations pass first guess quality control checks (This is what the red caption in the plot refers to). This is an indication that the surface pressure model error correction is particularly useful in the Ocean, where the observing system is significantly sparser than over land and thus model errors play a bigger role in the forecast error budget.

4.4 Evaluation of forecast skill

Here we concentrate on two aspects of the forecast performance of the ANN driven WC-4DVar experiments. The first aspect is whether they are able to replicate the improvements in stratospheric temperature reductions of forecast bias produced by the recent version of WC-4DVar (Laloyaux, Bonavita, Dahoui, et al., 2020). 10-day forecasts are initialised using the analysis from strong-constraint 4D-Var, weak-constraint 4D-Var, NN_SC and NN_WC between 10 and 24 August 2019. The model used to compute these forecasts is not corrected by any forcing estimated in weak-constraint 4D-Var or neural networks. Given the possible problems of correlated analysis and forecast errors that a standard own-analysis verification is likely to cause, we present forecast verification results against independent GPS-RO derived temperature profiles. Figure 12 shows the difference in temperature forecast RMSE after 72 hours between forecasts initialised by NN_SC and strong-constraint 4D-Var (a) and by NN_WC and strong constraint 4D-Var (b) and by weak-constraint 4D-Var and strong-constraint 4D-Var (c). The improvements obtained by weak-constraint 4D-Var are replicated by the two neural networks to a large extent. Degradations observed

at different pressure levels and latitudes are mainly not statistically significant. Comparing the two neural network approaches, one can see that weak-constraint 4D-Var used in NN_WC mitigates the degradation observed in NN_SC. Longer experiments are currently running to improve the statistical robustness of these results.

The other main question that we would like to answer is whether the introduction of model error forcing in the troposphere in the ANN driven 4DVar experiments is beneficial or not in terms of synoptic performance of the forecast. This is of particular interest in light of the fact that previous attempts to extend the current WC-4DVar formulation to the full atmospheric column resulted in significant degradations of various aspects of tropospheric forecast skill for the reasons explained in Laloyaux, Bonavita, Dahoui, et al. (2020). In Figure 13, we present two standard measures of synoptic performance for the mass field. Both 500 hPa geopotential (a-b) and Mean Sea Level Pressure (c-e) forecast errors for either of the ANN configurations appear slightly better than the reference strong and weak constraint 4D-Var, though statistical significance is only reached sporadically in the relatively short test period used here. Forecast performance for the wind field (not shown) is similar to that of the mass field presented earlier: No significant degradation is apparent, and some localised improvements consistent with the positive indications coming from the observation space assimilation diagnostics presented in Section 4.1 and 4.2 are also visible.

5 Discussion and research perspectives

The work presented in this paper and recent developments in 4DVar methodology (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, & Gürol, 2020) are based on the idea that an effective strategy to deal with model error in NWP is to partition it in two components: a) a stochastic, small scale (temporally and spatially) component and b) a predictable component active on larger and longer spatial/temporal scales. The random component of model error is typically represented with physically-based model error simulation models (Leutbecher et al., 2017) which are derived from an understanding of the approximations done in the development of the forecast model and an attempt to sample from those sources of uncertainties. These stochastic models of model error are then applied both in an ensemble data assimilation framework (Bonavita et al., 2012; Bowler, 2017) and ensemble forecast mode (Leutbecher et al., 2017). In ensemble data assimilation, their net effect is to produce a flow-dependent increase of ensemble spread and an associated improvement in the ensemble forecast reliability budget, which is usually under-dispersive (Houtekamer & Zhang, 2016; Bonavita et al., 2012; Bowler, 2017). These model error parameterisations are targeted at improving the ensemble estimate of the second order moment of the forecast error pdf. They might also be able to indirectly affect the ensemble forecast mean due to nonlinear effects arising during the model integrations, but these effects are small in a data assimilation cycling context (and often explicitly discarded through re-centring techniques). The second component of model error, which we have considered in this work, is the large scale error that evolves slowly over the time scale of the assimilation window length. We posit that this error is predictable, i.e. we can estimate the first moment of its distribution through statistical estimation techniques. Weak constraint 4DVar is an on-line example of a statistical estimation technique for dealing with the systematic errors of the model. The machine learning models described in this paper are examples of off-line statistical models aimed at achieving similar goals. A variety of hybrid configurations with a combination of on-line and off-line estimators are also possible: the WC-4DVar configuration where the ANN model error estimate is used as first guess and background for the WC-4DVar minimisation (NN_WC) is just an initial, proof-of-concept attempt. Similarly to the stochastic model error parameterisations, the use of WC-4DVar or its ML hybrids can improve reliability in an ensemble assimilation and forecasting system, but through

a different mechanism, namely reducing the total error budget by reducing/removing the systematic error components.

5.1 Research perspectives

The preliminary results presented in the previous section show that combining ANN models and WC-4DVar holds promise of improving on each technique used in isolation. In particular, it appears that a hybrid ANN–WC-4Dvar setup can be configured to effectively reduce model error throughout the atmospheric column and not only in the stratosphere. The specific configuration of this hybrid ANN-WC-4DVar is being currently investigated and the findings of this research will be reported in a follow-up paper. Other aspects of the methodology presented in this work can be further improved. Of fundamental importance is the choice of predictands and predictors for the model error regression problem. In terms of model error predictands, we have chosen to use analysis increments in state space. This idea is not new in NWP (Dee, 2005) and stems from the somewhat obvious consideration that only observations can (directly or indirectly) tell us something useful about model error with respect to the real atmosphere. This idea has been more recently revived in an ensemble data assimilation and forecasting context by Bowler (2017), following earlier work by (Piccolo & Cullen, 2015). With respect to these later works, our application differs in two important aspects: a) it can also be applied in a deterministic, perfect model assimilation and forecasting system, and b) its estimates are derived from flow-dependent regressions. This second property is important not only because the flow-dependent component adds more predictive power than that coming from climatological predictors (Section 2), but because it opens the perspective of using ANN models as an online, flow-dependent model error correction forcing term. This will be potentially interesting for improving predictions at longer forecast ranges than those considered in this work, as the accuracy and reliability of ensemble forecast predictions at long forecast scales are notoriously affected by the model systematic errors. At the current stage of research it is unclear whether this application of ANN, WC-4DVar or their hybrids will be practically successful. This is also in view of the complex and typically non-linear model error interactions that arise between the various components of a coupled Earth System model during extended integrations. We note however, that recent results in both medium range NWP (Laloyaux, Bonavita, Dahoui, et al., 2020) and seasonal prediction (Ham et al., 2019) have already shown that the introduction of pure or hybrid ML/DL models can lead to significant improvements in specific aspects of forecast performance.

The choice of analysis increments as model error predictands was mainly dictated by reasons of convenience and practicality. Another option is to directly use observation departures. This would have the advantage of avoiding another source of errors from the data assimilation system. On the other hand, one is limited to the relatively small subset of observations which are thought not to be affected by significant systematic errors themselves (e.g., radiosondes, GPS-RO), issues connected with their spatial and temporal homogeneity and more complex relations of the observed to the state variables. Still, we believe these issues can be addressed to some extent and a separate research effort is ongoing in this direction at ECMWF. Another area of development regards the type and choice of predictors used for our regression problem. As shown in Section 2.2 a judicious choice of additional predictors can further improve the predictive power of the ANN model and, likely, its impact on the IFS analyses and forecasts. Additional predictors can be envisaged which exploit additional sources of predictability of the atmospheric flow, especially those coming from fixed or slow-evolving boundary conditions (e.g., orography, land use, sea surface temperatures, etc.). In the context of choosing appropriate sets of predictors and predictands, their geometry and spatial resolution, the issues connected to overfitting and the so-

called "curse of dimensionality" become prominent. We discuss them in the following sub-section.

5.2 Statistical regression and the curse of dimensionality

As a type of statistical learning, machine learning is exposed to the problem of the "curse of dimensionality". Loosely speaking, this means that for systems with a large number of degrees of freedom, the number of available training examples will always be much smaller than the number of possible configurations in state space. Standard results from ergodic theory of dynamical systems (Cecconi et al., 2012) show that, for a dissipative nonlinear dynamical system like the atmosphere (or a state of the art NWP model), the minimum length M of the time series of past observed states (i.e., the size of the training dataset) necessary to find an analogue of the current state within an error distance measure ϵ has a scaling law of the form:

$$M \sim \left(\frac{L}{\epsilon}\right)^{D_a} \tag{10}$$

where L represents a measure of the variability of the system and D_a is the effective dimension of the system attractor, which can be a non integer number (i.e., a strange attractor). The exponential dependence of the training dataset size on the effective attractor dimension makes a fully statistical approach to forecasting unfeasible (Van den Dool, 1994). For machine learning applications to NWP and climate they indicate that an acritical application of ML tools is not likely to give good results unless effective mitigating strategies are put in place. There are at least two possible avenues to combat the curse of dimensionality. One is obviously to try to reduce the size of the regression or classification problem. This has motivated our choice to deal with the model error estimation problem, which can be framed as the problem of trying to identify a residual model which fills the gap between the actual forecast model and reality (cp. Eq. 2). It is reasonable to assume that the attractor dimension of the residual model is much smaller than that of the full model, as suggested, for example, in studies using reduced order models (Watson, 2019). The applicability of this assumption is further strengthened by the choice of using a coarse spatial resolution for the training dataset. This limits the modes of variability allowed in the regression and allows to train the ML model on a relatively small dataset and achieve good generalisation performance. The other standard tool to beat the curse of dimensionality is to use prior knowledge about the data generating distribution to suitably restrict the choice of the model space in which the machine learning algorithm is allowed to search for solutions citep(this is called the "hypothesis space" in machine learning literature). This is where expert knowledge of the problem at hand becomes valuable as there is no machine learning algorithm that is universally better on all possible tasks (The so-called "No free lunch" theorem Wolpert, 1996). With this in mind, we have chosen the avenue of training a fully connected ANN over atmospheric columns of predictor and predictand examples. The insight here is that it is important for the regression model to learn vertically balanced increments to avoid introducing spurious unphysical instabilities in the model evolution. This is also consistent with the way standard NWP and Climate Prediction models are currently formulated: the equations governing the model physical tendencies are typically formulated over model columns. Other approaches are possible, e.g. use Convolutional Neural Networks of the type that are currently popular in image recognition applications to try to learn spatial patterns on model levels. We leave this for future investigation. An additional advantage of our choice of predictors and predictands geometry is that it helps to drastically reduce the number of learnable parameters of the ANN model and thus the risk of overfitting the training dataset.

6 Conclusions

Machine Learning and Deep Learning technologies have been applied successfully in many disparate fields. These remarkable success stories have in the past few years generated interest in the NWP and climate communities to understand whether there is scope to apply ML/DL techniques in their respective fields. However, while a number of visionary, speculative papers have been published explaining the case for the application of ML/DL to NWP and climate, and an even greater number have investigated the use of ML/DL techniques in a variety of low-order models, very little work seems yet to have been undertaken to apply ML/DL methods to state-of-the-art, high resolution global circulation models such as those used in operational global NWP and climate. The work presented here aims at starting to fill this gap. The results presented in this paper show a first application of ML/DL tools to the problem of model error estimation and correction in a data assimilation context. Building on recent results obtained in a weak constraint 4D-Var framework (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, & Gürol, 2020) we show that the use of ANN-derived model error forecasts potentially allows to extend the benefits of the weak constraint formulation of 4D-Var to the troposphere, which had been an intractable problem since the introduction of WC-4DVar at ECMWF more than ten years ago. While these results need to be validated over longer testing periods and the technical infrastructure is not yet fully in place for reliable operational use, we believe these results to be promising enough to warrant actively pursuing this line of research further.

From the vantage point of data assimilation in the geosciences, ML/DL do not introduce completely new or revolutionary ideas. In fact, ML/DL techniques and their theoretical underpinnings have much in common with the standard toolbox of variational data assimilation, though these similarities are partly obfuscated by the different nomenclature (Geer, 2020). What the most recent ML/DL wave of interest has brought about is the availability of a set of powerful, easy to use, open source libraries which greatly facilitate the application of these techniques to disparate fields; and a renewed awareness of the effectiveness of these techniques in many different contexts. These newly available tools are undoubtedly helping the adoption of ML/DL techniques in the NWP community beyond already well-established areas such as e.g. NWP output post-processing and nowcasting (McGovern et al., 2017; Gagne et al., 2017; Rasp & Lerch, 2018). As discussed in Section 5, enthusiasm for adopting these new tools in core NWP tasks needs to be tempered by a careful appreciation of their fundamental limitations. Even with the huge increase in modern computational resources and the size of available training datasets, it is unlikely, for example, that a fully ML/DL-based forecast model will supersede the current knowledge-based forecast models. On the other hand, it is possible and even probable that knowledge-based models of the not-too-distant future will integrate ML/DL components for reasons of computational efficiency and possibly improved performance. At the same time, it is likely that ML/DL tools for model error estimation and correction like those presented in this work will play a major role in a variety of data assimilation and forecasting applications.

Acknowledgments

We gratefully acknowledge constructive discussions with a number of colleagues at ECMWF and beyond. In particular, Marc Bocquet, Alban Farchi, Peter Dueben, Alan Geer, Peter Bauer, Nils Wedi and Elias Hólm have reviewed an earlier version of the manuscript and provided many useful suggestions for its improvement. We also wish to express our gratitude to our ECMWF colleague Xavier Abellan for his patient and precious technical support during the course of this project. Data availability statement: The input and output data of the experiments described in the paper is

freely available for research purposes from ECMWF and can be requested following the procedures described in https://www.ecmwf.int/en/forecasts/datasets

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th usenix symposium on operating systems design and implementation (osdi 16)* (pp. 265–283). Retrieved from `https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf`

Auligné, T., McNally, A. P., & Dee, D. P. (2007). Adaptive bias correction for satellite data in a numerical weather prediction system. *Quarterly Journal of the Royal Meteorological Society*, *133*(624), 631-642.

Bocquet, M., Brajard, J., Carrassi, A., & Bertino, L. (2020). Bayesian inference of chaotic dynamics by merging data assimilation, machine learning and expectation-maximization. *Foundations of Data Science*, *2*, 55–80.

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376-399.

Bonavita, M., Isaksen, L., & Holm, E. (2012). On the use of EDA background error variances in the ecmwf 4D-Var. *Quarterly Journal of the Royal Meteorological Society*, *138*, 1540–1559.

Bonavita, M., Lean, P., & Holm, E. (2018). Nonlinear effects in 4d-var. *Nonlinear Processes in Geophysics*, *25*(3), 713–729.

Bowler, N. E. (2017). On the diagnosis of model error statistics using weak-constraint data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *143*(705), 1916-1928.

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the lorenz 96 model. *Journal of Computational Science*, 101–171.

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887-2908.

Cecconi, F. M., Cencini, M., Falcioni, & Vulpiani, A. (2012). The prediction of future from the past: an old problem from a modern perspective. *American Journal of Physics*, *80*(11), 1001-1008.

Chollet, F. (2018). *Deep learning with python*. Manning Publicationa, New-York.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Control Signal Systems*, *2*, 303–314.

Dee, D. P. (2005). Bias and data assimilation. *Quarterly Journal of the Royal Meteorological Society*, *131*(613), 3323–3343.

Gagne, I., David John, McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017, 09). Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, *32*(5), 1819-1840.

Geer, A. (2020). Learning earth system models from observations: machine learning or data assimilation? *Philosophical Transactions of the Royal Society*, Submitted. doi: 10.21957/7fyj2811r

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press, Cambridge.

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019, 09). Deep learning for multi-year enso forecasts. *Nature*, *573*. doi: 10.1038/s41586-019-1559-7

Hamrud, M., Bonavita, M., & Isaksen, L. (2015, 11). EnKF and Hybrid Gain En-

semble Data Assimilation. Part I: EnKF Implementation. *Monthly Weather Review*, *143*(12), 4847-4864.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. doi: 10.1002/qj.3803

Houtekamer, P. L., & Zhang, F. (2016). Review of the Ensemble Kalman Filter for Atmospheric Data Assimilation. *Monthly Weather Review*, *144*(12), 4489-4532.

Jaeger, H. (2001). *The "echo state" approach to analysing and training recurrent neural networks-with an erratum note* (GMD Technical Report No. 148). German National Research Center for Information Technology.

Kalman, R. E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, *82*(1), 35-45.

Kingma, D., & Ba, J. (2014, 12). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Laloyaux, P., Bonavita, M., Chrust, M., & Gürol, S. (2020). Exploring the potentials and limitations of weak-constraint 4d-var. *Quarterly Journal of the Royal Meteorological Society*.

Laloyaux, P., Bonavita, M., Dahoui, M., Farnan, J., Healy, S., Hólm, E., & Lang, S. T. K. (2020). Towards an unbiased stratospheric analysis. *Quarterly Journal of the Royal Meteorological Society*, https://doi.org/10.1002/qj.3798.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., ... 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*(707), 2315-2339.

McGovern, A., Elmore, K. L., Gagne, I., David John, Haupt, S. E., Karstens, C. D., Lagerquist, R., ... Williams, J. K. (2017, 10). Using Artificial Intelligence to Improve Real-Time Decision-Making for High-Impact Weather. *Bulletin of the American Meteorological Society*, *98*(10), 2073-2090.

Pathak, J., Wikner, A., Fussell, R., Chandra, S., Hunt, B. R., Girvan, M., & Ott, E. (2018). Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(4), 041101. doi: 10.1063/1.5028373

Piccolo, C., & Cullen, M. (2015, 12). Ensemble Data Assimilation Using a Unified Representation of Model Error. *Monthly Weather Review*, *144*(1), 213-224.

Raanes, P. N., Bocquet, M., & Carrassi, A. (2019). Adaptive covariance inflation in the ensemble kalman filter by gaussian scale mixtures. *Quarterly Journal of the Royal Meteorological Society*, *145*(718), 53-75.

Rasp, S., & Lerch, S. (2018, 10). Neural Networks for Postprocessing Ensemble Weather Forecasts. *Monthly Weather Review*, *146*(11), 3885-3900.

Rodwell, M., & Jung, T. (2008, 07). Understanding the local and global impacts of model physics changes: An aerosol example. *Quarterly Journal of the Royal Meteorological Society*, *134*, 1479 - 1497. doi: 10.1002/qj.298

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014, January). Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, *15*(1), 1929–1958.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Tremolet, Y. (2006). Accounting for an imperfect model in 4d-var. *Quarterly Journal of the Royal Meteorological Society*, *132*(621), 2483-2504.

Van den Dool, H. M. (1994). Searching for analogues, how long must we wait? *Tellus A*(46), 314-324.

Watson, P. A. G. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. *Journal of Advances in Modeling Earth Systems*, *11*(5), 1402-1417.

Whitaker, J. S., & Hamill, T. M. (2012). Evaluating Methods to Account for System Errors in Ensemble Data Assimilation. *Monthly Weather Review*, *140*(9), 3078-3089.

Wolpert, D. H. (1996). The lack of a priori distinctions between learning algorithms. *Neural Computation*, *8*(7), 1341-1390.

Figure.

853 **List of Figures**

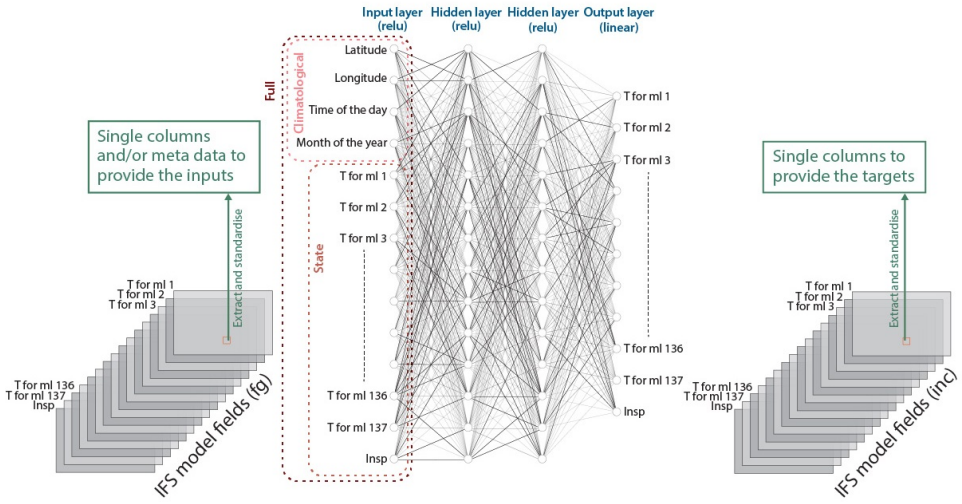

Figure 1. Diagram representing how the relu_three_layers ANN is built for the regression over temperature and logarithm of surface pressure. Single columns plus metadata (latitude, longitude, time of the day and month of the year) are extracted from the first guess and analysis increment gridded fields to produce the input and the target of the neural network. Climatological neural network uses only the metadata as input while state neural network uses only the temperature and logarithm of surface pressure values as input. The full neural network combines both information. All neural networks used in this work contain a certain number of fully connected nonlinear layers and an output linear layer.

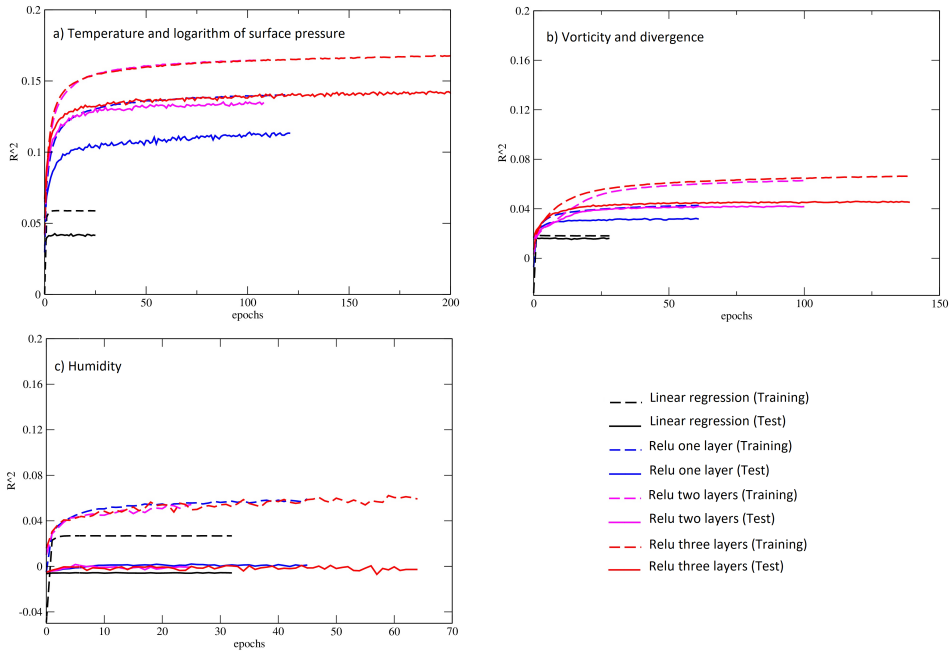

Figure 2. Evolution of the R2 coefficient against the number of epochs during the training of different ANN for the (t, lnsp) predictands (a), for the wind (vo, d) error set of predictands (b) and for the humidity (q) error set of predictands (c). Dashed lines refer to R2 values over the training dataset, continuous lines to R2 values over the test dataset.

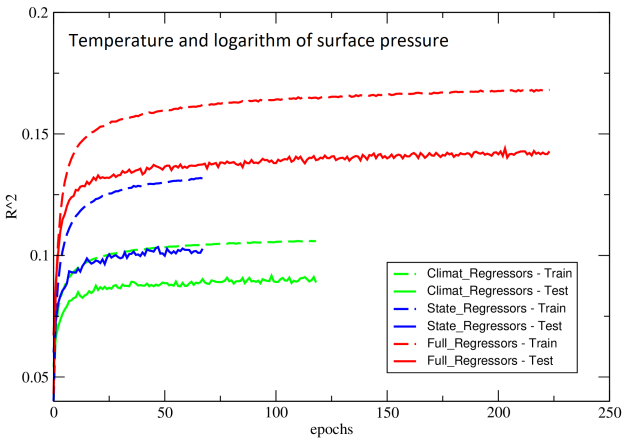

Figure 3. R2 coefficients computed during the training phase of the relu_three_layers ANNs using only climatological predictors (green curves); only state predictors (blue curves); both sets of predictors (red curves). Dashed curves show R2 coefficients for the training dataset, continuous curves for the test dataset.

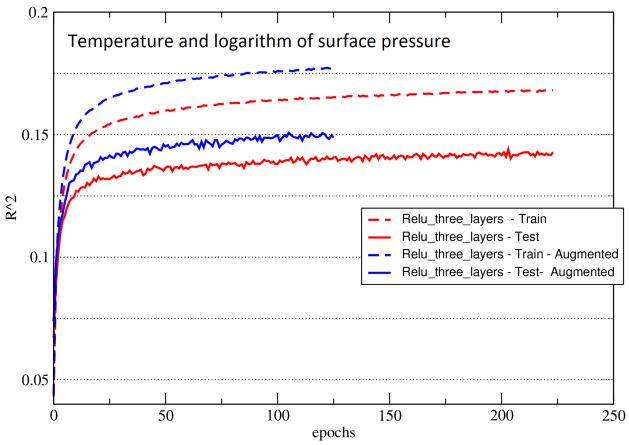

Figure 4. R2 coefficients computed during the training phase of a relu_three_layers ANN for the prediction of (t, lnsp) model errors using the standard set of predictors (red lines) and the augmented set (blue lines). Dashed lines for the training dataset, continuous lines for the test dataset.

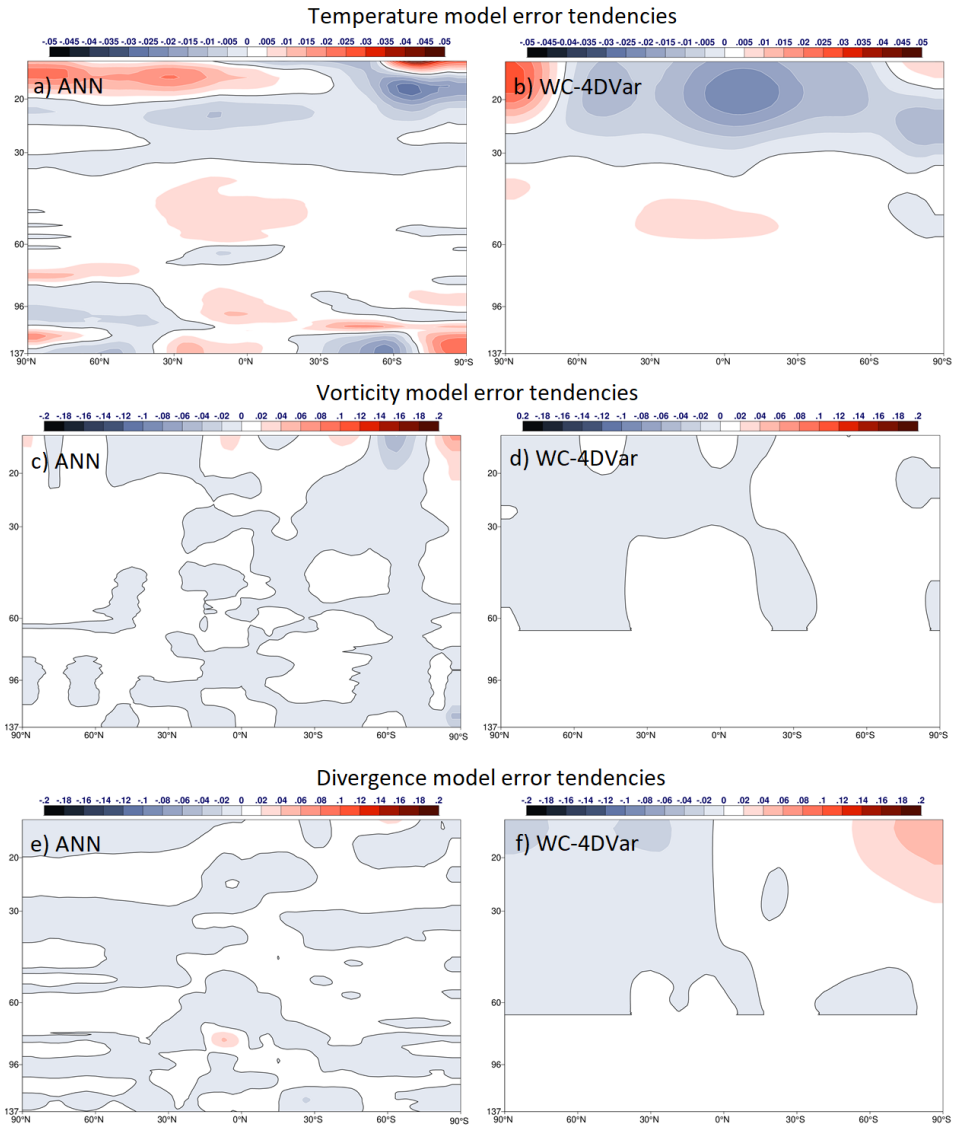

Figure 5. Vertical profiles from model level 15 (approx. 1 hPa) to model level 137 (bottom model level) of longitudinal average of the temperature (a-b), vorticity (c-d) and divergence (e-f) model error tendencies estimated by the ANN described in the text (left panels) and by WC-4DVar from a pre-operational version of IFS cycle 47R1 (right panels). IFS model levels (20, 30, 60, 96, 137) correspond approx. to pressure levels (2 hPa, 10 hPa, 100 hPa, 500 hPa, surface). Values are averaged over a one week period starting on 2019-07-20. The thick black line separates negative tendencies (shades of blue) to positive tendencies (shades of red).

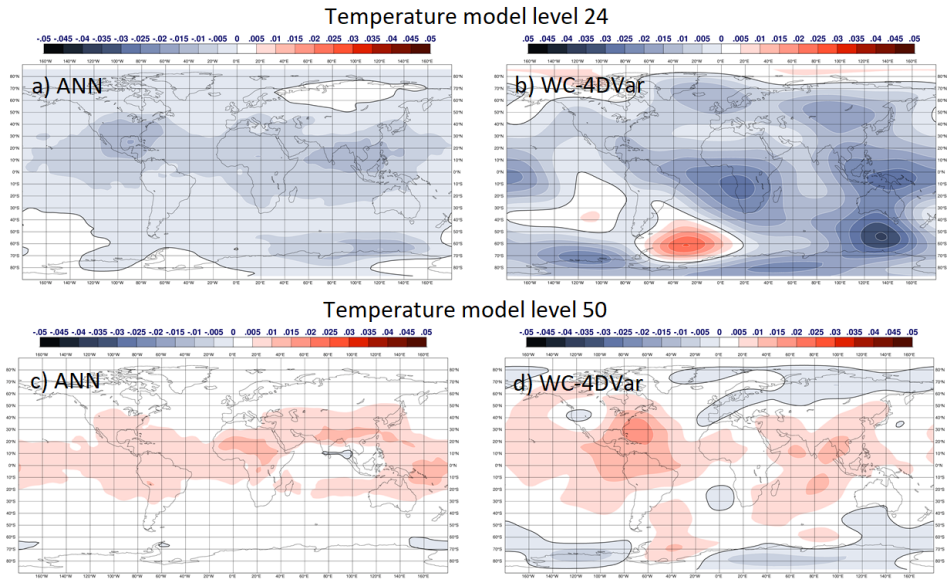

Figure 6. Geographical maps at model level 24, approx. 5 hPa (a-b) and model level 50, approx. 50 hPa (c-d) of the temperature model error tendencies estimated by the ANN (left panels) and by WC-4DVar from a pre-operational version of IFS cycle 47R1 (right panels). Values are averaged over a one week period starting on 2019-07-20. Units in kelvin/h. The thick black line separates negative tendencies (shades of blue) to positive tendencies (shades of red)

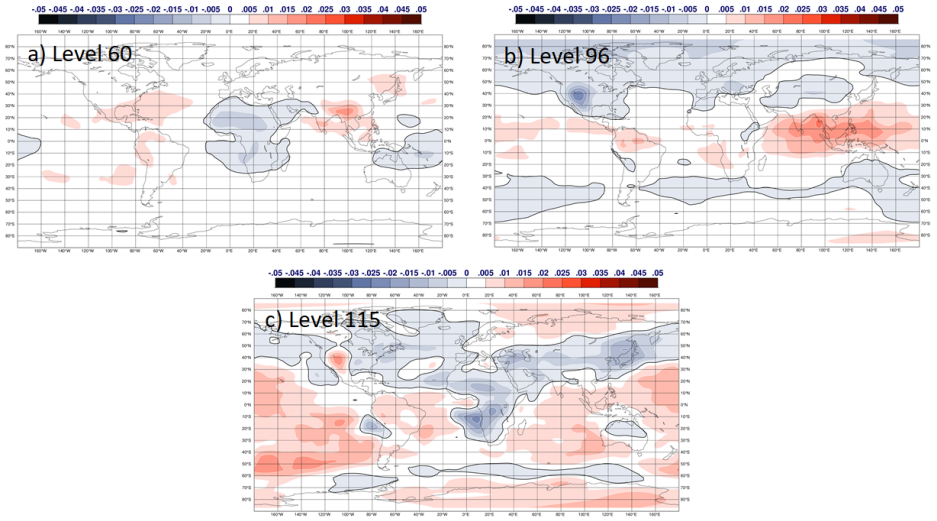

Figure 7. As in Figure 6 for model level 60, approx. 100 hPa (a), model level 96, approx. 500 hPa (b) and model level 115, approx. 850 hPa (c).

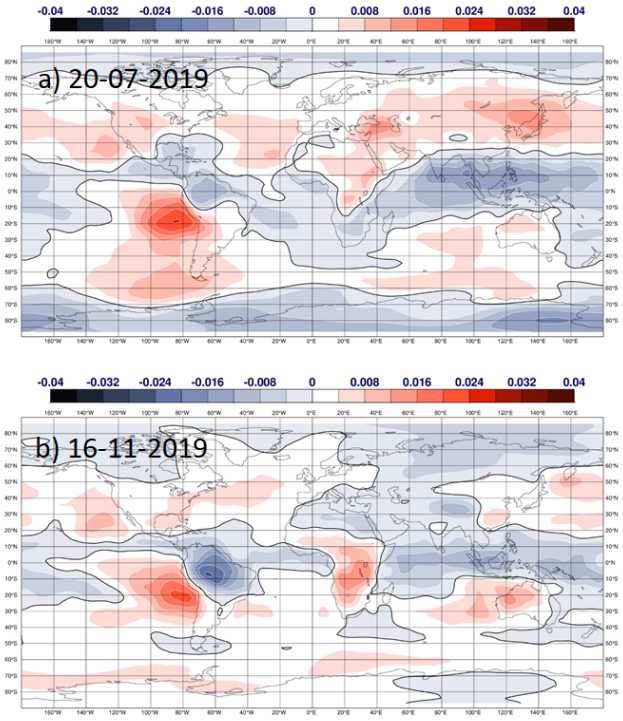

Figure 8. Geographical maps of the surface pressure model error tendencies estimated by the ANN. Values are averaged over two one-week periods starting on 2019-07-20 (a) and on 2019-11-16 (b). Units in hPa/hour. The thick black line separates negative tendencies (shades of blue) to positive tendencies (shades of red)

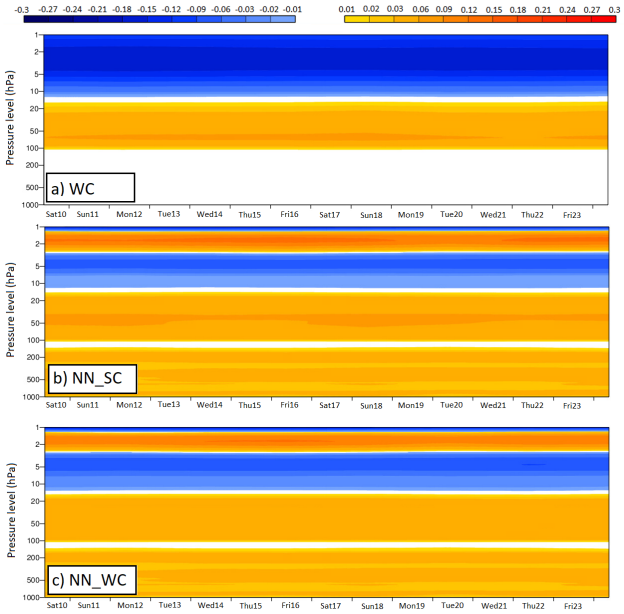

Figure 9. Time series of the global mean model error correction estimated with weak-constraint 4D-Var (top), NN_SC (middle) and NN_WC (bottom) from 15 to 24 August 2019.

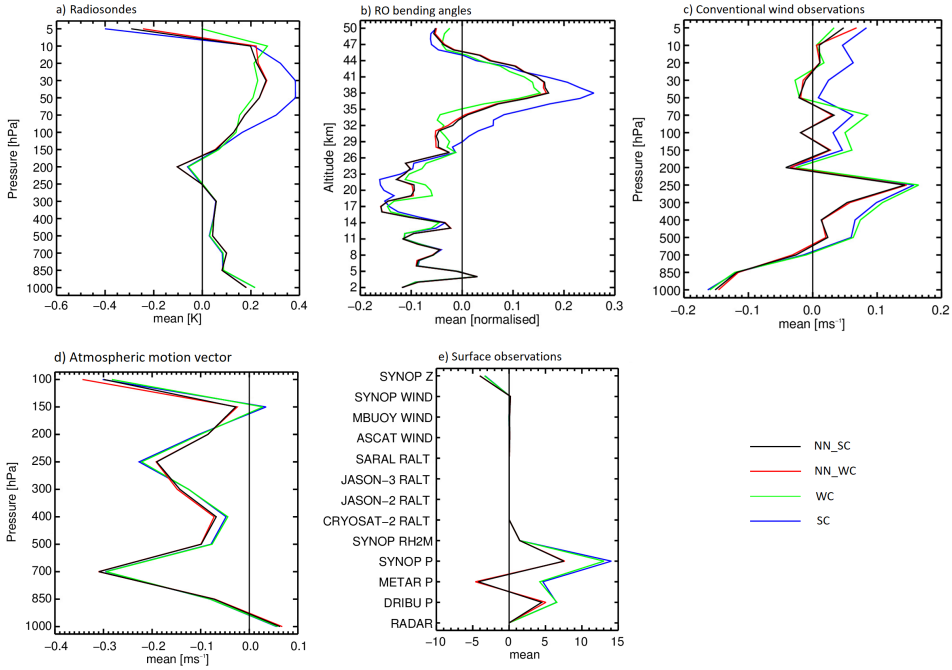

Figure 10. Mean background departures for radiosonde temperature (a), radio occultation bending angles (b), conventional wind observations (c), atmospheric motion vectors (d) and surface observations (e) for the four experiments described in the main text (SC: blue; WC: green; NN_SC: black; NN_WC: red). Values averaged over the 16-07-2019 to 24-08-2019 period. The legend entries SYNOP P, METAR P and DRIBU P stand for surface pressure observations from Synop stations, Metar stations and Drifting Buoys.

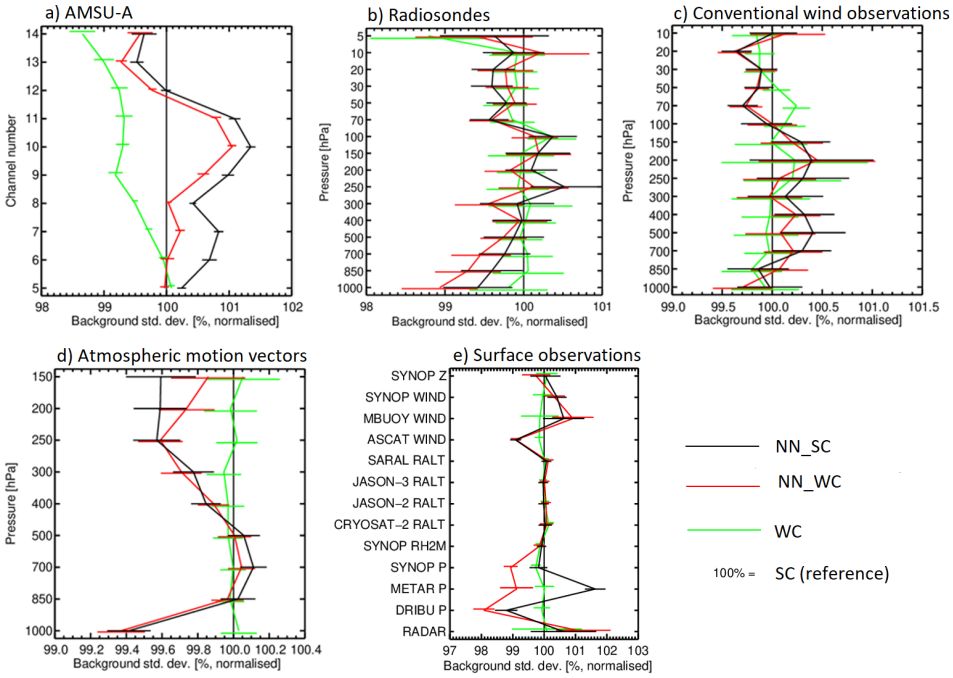

Figure 11. Normalised standard deviation of background departures and background departures for AMSU-A radiances (a), radiosonde temperatures (b), conventional wind observations (c), atmospheric motion vectors (d) and surface observations (e) for the four experiments described in the main text (SC: reference 100%; WC: green; NN_SC: black; NN_WC: red). Values averaged over the 16-07-2019 to 24-08-2019 period. The legend entries SYNOP P, METAR P and DRIBU P stand for surface pressure observations from Synop stations, Metar stations and Drifting Buoys).

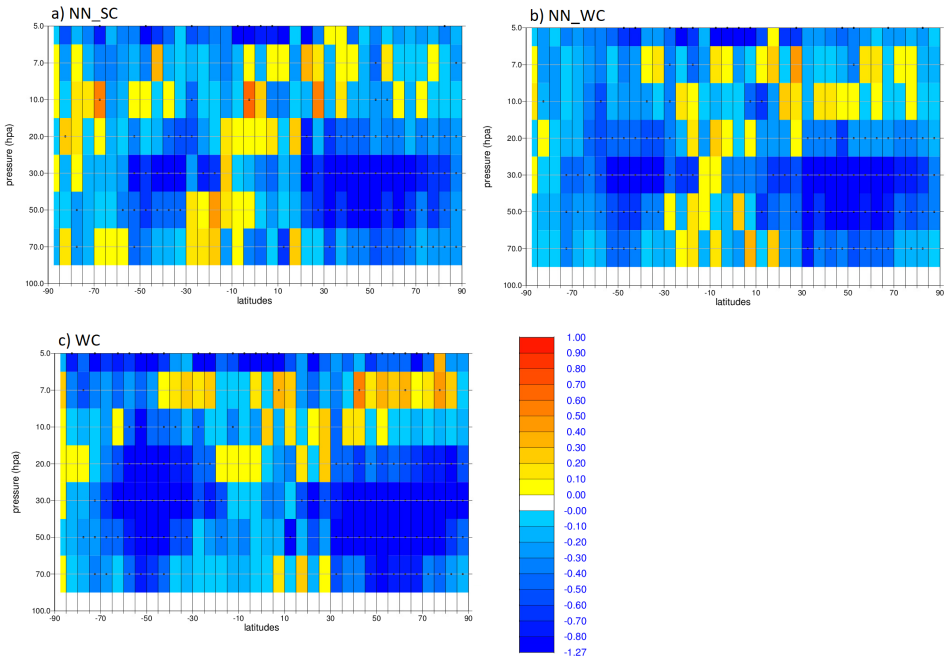

Figure 12. Difference in temperature forecast RMSE after 72 hours between forecasts initialised by NN_SC and strong-constraint 4D-Var (a) and by NN_WC and strong constraint 4D-Var (b) and by weak-constraint 4D-Var and strong-constraint 4D-Var (c). RMSE is computed using radio occultation temperature retrievals and averaged between 10 and 24 August 2019. A negative (positive) difference means that the new system reduces (increases) the forecast error with respect to strong-constraint 4D-Var. Black dots indicate areas where the signal is significant at the 95% confidence level.

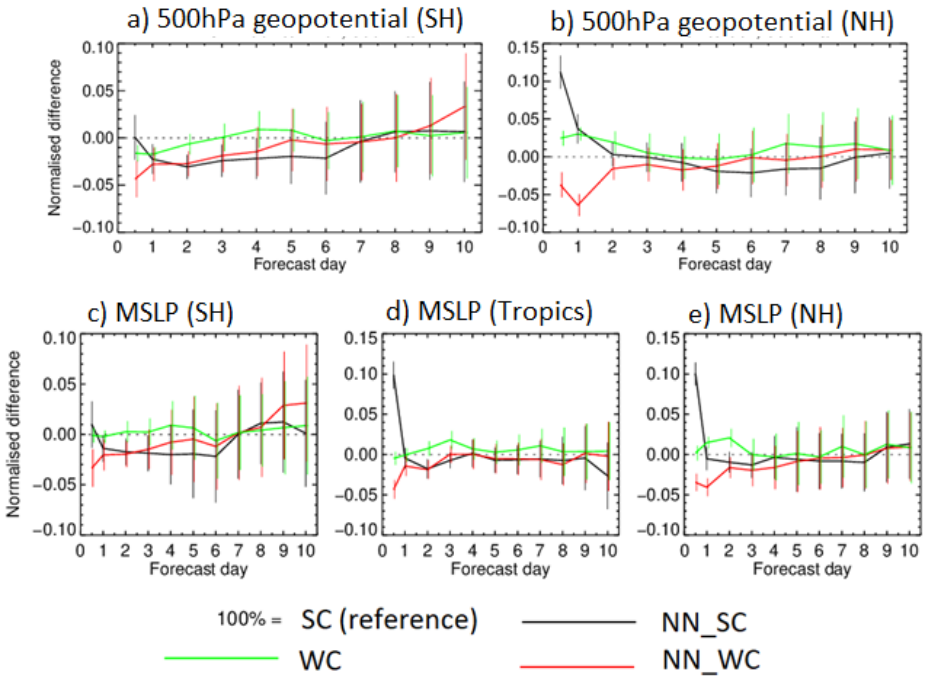

Figure 13. Left panel: Normalised forecast root mean square error of the 500 hPa geopotential in the southern (a) and northern (b) hemisphere, as well as normalised forecast root mean square error of the mean see level pressure field in the southern hemisphere (c), tropics (d) and northern hemisphere (e). Reference zero line is the strong constraint 4DVar experiment error level. Values averaged over the 16-07-2019 to 24-08-2019 period. Verification is against own analysis. Error bars denote 95% confidence levels.