Machine Learning for Model Error Inference and Correction

Massimo Bonavita and Patrick Laloyaux

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

Abstract Model error is one of the main obstacles to improved accuracy and reliability in numerical weather prediction (NWP) and climate prediction conducted with state-of-the-art, 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 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 (ANNs) can reproduce the main results obtained with weak-constraint 4D-Var in the operational configuration of the IFS model of the European Centre for Medium-Range Weather Forecasts (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 the European Centre for Medium-Range Weather Forecasts (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 here 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.

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, formalized in the evolution equation of the state error covariance in the Kalman filter (Kalman, 1960)

\[ P_{t} = M P_{t-1} M^{T} + Q_{t}, \]

where the error covariance matrix of the background forecast state \( P_{t} \) is written as the sum of the evolved analysis errors from the previous analysis update \( (MP_{t-1}M^{T}) \), where \( M \) is the linear/linearized model and a zero-mean stochastic model error of covariance \( Q_{t} \). 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 to some extent
relaxed in operational settings. For example, at the European Centre for Medium-Range Weather Forecasts (ECMWF) the model error parameterizations used in the Ensemble of Data Assimilations (EDA) to simulate model error evolution are based on a multiplicative ansatz (Buizza et al., 1999), and the spatial structures of 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). The term bias is used here with a somewhat wider meaning than that usually associated with it 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, and in different meteorological conditions, and they can be also influenced by systematic errors arising from the interactions 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. Recognizing 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, suboptimal 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, and 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, suboptimal 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 estimate of the model error forcing (this is the analyzed model error from the previous WC-4DVar; i.e., the identity operator from one assimilation window to the next), and $Q$ 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, which became operational in 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 the background error $B$ and of the model systematic errors $Q$ (see Laloyaux, Bonavita, Chrust, et al., 2020; Laloyaux, Bonavita, Dahoui, et al., 2020, for a detailed explanation). The scale separation allows to successfully de-alias initial state and model error corrections during the 4D-Var minimization and is consistent with the 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 online machine learning (ML) algorithm.

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 unrivaled 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 they aim to learn from a data set of examples $(X, Y)$, a (possibly) nonlinear mapping between features $X = \{x_1, x_m\}$, and a corresponding set of observed targets (labels) $Y$. This is usually done by assuming a parametric model for the conditional distribution of the observations, $p_{\text{model}}(Y|X, \theta)$ and maximizing the likelihood of the model over the empirical data distribution $p_{\text{obs}}(Y|X)$

$$\theta_{\text{ML}} = \arg \max_{\theta} p_{\text{model}}(Y|X, \theta).$$

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

$$\theta_{\text{ML}} = \arg \max_{\theta} \sum_{i=1}^{m} \log \left( p_{\text{model}}(y_i|x_i, \theta) \right),$$

where $i$ is the index running over the $m$ members of the examples data set. This is equivalent to minimizing 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 $(Y, X)$ represented by the affine transformation

$$Y = WX + b.$$  

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

$$p_{\text{model}}(Y|X, \theta) = \mathcal{N}(Y|WX + b, I),$$

where the general set of learnable parameters $\theta$ has been particularized to the sets of weights $W$ and bias coefficients $b$ of a generic neural network. Maximizing the log likelihood (or, more commonly, minimizing the negative log likelihood) of this model leads to the standard normal equations. Adding constraints on the size of the regression coefficients matrix $W$ (known in different communities as Tikhonov regularization, ridge regression, weight decay) or the sparsity of said matrix (Lasso: Tibshirani, 1996) can be seen as ways of improving the generalization 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 $(Y, X)$ is nonlinear, then the maximum likelihood estimator in Equation 5 will be suboptimal. In our problem of model error estimation, it is a priori unclear how big 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 relinearization steps (Bonavita et al., 2018). In DL, 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 multilayer 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 data set.

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 hybridizing ML 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 and 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 reframed by Bocquet et al. (2020) into a unifying Bayesian formalism, 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 data assimilation and forecasting applications. These applications pose a set of additional challenges. First, 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). Second, and possibly more importantly, the dimensions of the analysis and forecast system in operational NWP are very large. In the current 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 $\mathcal{O}(10^5)$. 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 of successfully integrating 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 organized 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 current 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. ML Methodology and Results

2.1. Setup 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 $(X,Y)$ of the supervised learning problem). In a real-world setting we do not have access to the true model error predictands ($Y$); 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
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.

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 (Insp), vorticity (vo), divergence ($d$), and 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 time series 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., Equations 43 and 44 in Dee, 2005).

We can broadly consider two classes of predictors ($X$). 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 following 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, for example, oceanic stratocumulus areas, Intertropical Convergence Zone, and extratropical cycloic areas. 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 ($t$, Insp, 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.

The horizontal spatial resolution of the fields whose vertical columns are used in the regression (see Figure 1 for a schematic of the ANN structure) is an important design choice. For this work we have selected a resolution in spectral space of Triangular spectral truncation 21 (T21), which corresponds to performing a local average of the original T1279 fields to an approximate grid spacing of 900 km on a reduced quadratic Gaussian grid. This choice was motivated by both practical and fundamental considerations. 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) and Laloyaux, Bonavita, Chrust, et al. (2020) that only large-scale model errors are identifiable in a WC-4DVar 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 ANN was conducted on a dedicated dual GPU workstation (NVIDIA Quadro GV100) using the open source DL 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 speedup in the training phase achieved on the GPU system was a factor of approximately 3. The training data set consisted of operational analysis increments and background forecasts collected over the whole year of 2018 every 36 h (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 data set was composed in the same way using a short 2 month period on 1 January 2019. This data set was used to get an indication of appropriate hyperparameters values. The test data set used for verifying the performance of the ANN was composed of the analysis increments and background forecasts of a 3.5 month period starting on 1 April 2019.

The statistical regressions have been computed separately for the three sets 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 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 $142 \times 138 + 138 = 19,734$ (The size of the input vector comes from the concatenation of the first-guess column values, 137 + 1, and the four climatological predictors). The number of vertical profiles in the training data set is $\mathcal{O}(10^6)$, which, as we will see, is enough to prevent overfitting. 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 modeled by the REctified Linear Unit (RELU) function, expressed by the function

$$\text{Relu}(x) = \max(0, x).$$

(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 minimizer 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 hyperparameters (learning rate and decay rates) than standard SGD and other adaptive methods available in the Tensorflow toolbox. Regularization is also an important aspect of DL methodology. In our case we found regularization to be only moderately helpful, mainly improving monotonicity of convergence and slightly improving generalization 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 regularization methods is likely due to the relative shallowness of the ANN we have used and the fact that the size of our training data set is 1 to 2
orders of magnitude larger than the number of model parameters. The main hyperparameters of the ANN and their values are listed in Table 1 for quick reference.

As it is standard in regression settings, the mean square difference between predicted and actual model errors is minimized 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/test data sets that is explained by the model

\[
R^2 = 1 - \frac{SS_{res}}{SS_{tot}},
\]

where \(SS_{tot} = \sum_{i=1}^{m}(y_i - \bar{y})^2\) is the total sum of squares (proportional to the sample variance), and \(SS_{res} = \sum_{i=1}^{m}(y_i - \hat{y})^2\) is the residual sum of squares. In a perfect model scenario where the model is able to accurately predict every instance of the sampling data set, \(R^2 = 1\). As our model error generating processes are inherently stochastic, even a perfect model, that is, a model that makes predictions sampling from the true error generating distribution, will produce some error, so that \(R^2\) 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 \(\bar{y}\) has a \(R^2 = 0\), and models that do worse than this baseline will have negative \(R^2\). The \(R^2\) coefficient has also been used in this work as stopping criterion in the training to avoid overfitting (i.e., training is stopped when \(R^2\) has not increased over the previous 20 epochs). In Figure 2, we present the results of the ANN training for the three sets of model error tendency predictands: \((t, \ln sp)\), \((v, d)\), 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, and month of

---

**Table 1**

| Name                  | Options          | Parameter type |
|-----------------------|------------------|----------------|
| Batch normalization   | [yes, no]        | Choice         |
| Regularization        | [L2, Dropout]    | Choice         |
| Dropout               | [0, 0.2, 0.5]    | Continuous     |
| Activation function   | [Tanh, ReLU, SoftPlus] | Choice       |
| Number of layers      | [1–4]            | Discrete       |
| Optimizer             | [SGD, Adam, RMSProp] | Choice       |

---

**Figure 2.** Evolution of the \(R^2\) coefficient against the number of epochs during the training of different ANN for the \((t, \ln sp)\) predictands (a), for the wind \((v, d)\) error set of predictands (b), and for the humidity \((q)\) error set of predictands (c). Dashed lines refer to \(R^2\) values over the training data set, continuous lines to \(R^2\) values over the test data set.
Figure 3. $R^2$ coefficients computed during the training phase of the relu_three_layers ANNs using only climatological predictors (green curves), only state predictors (blue curves), and both sets of predictors (red curves). Dashed curves show $R^2$ coefficients for the training data set, continuous curves for the test data set.

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, Chrust, et al., 2020; Laloyaux, Bonavita, Dahoui, et al., 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 ($R^2_{state} \sim 10\%$, $R^2_{climat} \sim 8\%$), but both sets of predictors contribute independent information to the final regression model ($R^2_{full} \sim 14\%$).

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 data set would likely be required to obtain robust estimates of these small covariances. In a similar vein, the set of state predictors could be augmented to include any other state variables that could potentially covariate with the predictands, for example, vertical velocity, precipitation rate, and liquid water content. 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 h 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 flavor 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 ($R^2 \sim 15\%$ vs $14\%$). A similar improvement is seen for the wind error predictands ($R^2 \sim 6\%$ vs. $5\%$).
Figure 5. Vertical profiles from Model Level 15 (approximately 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 preoperational version of IFS cycle 47R1 (right panels). IFS model levels (20, 30, 60, 96, and 137) correspond approximately to pressure levels (2, 10, 100, and 500 hPa, and surface). Values are averaged over a 1 week period starting on 20 July 2019. The thick black line separates negative tendencies (shades of blue) to positive tendencies (shades of red).

while no improvement is visible in the humidity ANN results (not shown). How much these improvements in the predictive power of the ANN translate into actual analysis and forecast improvements has not been investigated in this work, and it is deferred to future work where the impact of this and other changes to the set of predictors will be analyzed in a comprehensive way.

3. Predicting Model Error With ANNs

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 1 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 WC-4DVar model error estimates for an experiment run over the same period and initialized 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 (approximately 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 Levels 60 and 30 (approximately 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. The two estimates show however large differences in the horizontal distribution of the errors in the top model layers, which we believe to be due to the more intermittent and unbalanced nature of the dynamics in this upper model layer and also to the effects of a very active sponge layer parameterization (Shepherd et al., 2018). In the vorticity and divergence model error plots (Figures 5c–5f) 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 parameterized convection scheme, the data assimilation system, or both.
To obtain further insight into 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 Levels 24 (approximately 5 hPa) and 50 (approximately 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 purely online estimates (a running mean model error estimate is updated with the latest batch of observations) and thus more sensitive to existing flow conditions than the ANN estimates, which provide model error estimates typical for those atmospheric conditions.

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 hPa (a), 500 hPa (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 seabords of continental landmasses (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) and 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 and convective areas like the maritime continent and the Amazons), 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 and 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
Figure 8. Geographical maps of the surface pressure model error tendencies estimated by the ANN. Values are averaged over two 1 week periods starting on 20 July 2019 (a) and on 16 November 2019 (b). Units in hPa/h. The thick black line separates negative tendencies (shades of blue) to positive tendencies (shades of red).

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 investigation.

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 minimization 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 h integration from one analysis update to the next. By dividing this quantity by the number of time steps used in the 12 h 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 off-line 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).

**Algorithm 1 NN_SC**

Loop over the assimilation cycles from 16 July 2019 to 24 August 2019:

Loop over the physical variables \( (t, \text{Inspr}, \text{vo}, d, q) \):

- Compute the model error forcing with the trained ANN at valid times.
- Concatenate the outputs in a vector \( \eta^b \).
- Minimize the strong constrain 4D-Var

\[
J_{SC}(x_0) = \frac{1}{2} (x_0 - x_0^b)^T R^{-1} (x_0 - x_0^b) \\
+ \frac{1}{2} \sum_{k=0}^{N} (\mathcal{H}_k(x_k) - y_k)^T R_k^{-1} (\mathcal{H}_k(x_k) - y_k)
\]

where the model trajectories are computed as

\[
x_k = \mathcal{M}_{k,k-1}(x_{k-1}) + \eta^b \quad \text{for } k = 1, \ldots, N
\]

The second option, called NN_WC, is based on the WC-4DVar presented in Equations 2 and 3 where the model forcing \( \eta \) is initialized by the ANN tendency (see Algorithm 2). The model forcing that comes out of the minimization of Equation 3 which contains the ANN tendency updated by the WC-4DVar minimization is not cycled. Instead, in the successive assimilation window the model forcing \( \eta^b \) used as background in the minimization is cold started from the corresponding ANN estimate valid at the start time of the assimilation window.

**Algorithm 2 NN_WC**

Loop over the assimilation cycles from 16 July 2019 to 24 August 2019:

Loop over the physical variables \( (t, \text{Inspr}, \text{vo}, d, q) \):

- Compute the model error forcing with the trained ANN at valid times.
- Concatenate the outputs in a vector \( \eta^b \).
- Minimize the weak-constraint 4D-Var

\[
J_{WC}(x_0, \eta) = \frac{1}{2} (x_0 - x_0^b)^T B^{-1} (x_0 - x_0^b) \\
+ \frac{1}{2} \sum_{k=0}^{N} (\mathcal{H}_k(x_k) - y_k)^T R_k^{-1} (\mathcal{H}_k(x_k) - y_k) \\
+ \frac{1}{2} (\eta - \eta^b)^T Q^{-1} (\eta - \eta^b)
\]

where the model trajectories are computed as

\[
x_k = \mathcal{M}_{k,k-1}(x_{k-1}) + \eta \quad \text{for } k = 1, \ldots, N
\]

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) and at the operational configuration for both the 4D-Var analysis step and the forecast step (TC01279 spectral truncation, approximately 9 km grid spacing), over the period 16 July 2019 to 24 August 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.

### 4.1. Time Evolution of the Model Error Estimates

Figure 9a shows the zonal cross section of the mean temperature model error correction estimated with WC-4DVar and averaged between 15 and 24 August 2019. The model error is initialized at the beginning of the experiment from operations and is then cycled over the 12 h assimilation windows. WC-4DVar 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 WC-4DVar. 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. On the other hand, WC-4DVar extrapolates the cold bias to the top of the model, due to the deep vertical structure of the prescribed model error covariance matrix.

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 diagnostics presented in Figure 10 aim at evaluating the ability of the two ANN-informed 4D-Var experiments to reduce mean short-range forecast errors in comparison to WC-4DVar. 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 behavior 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 approximately 700 hPa (Figure 10c). 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 WC-4DVar 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...
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 July to 24 August 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.

by showing in Figure 11 the normalized standard deviation of 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 (Channels 5 to 8) and the stratosphere (Channels 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 4D-Vars and equivalent to the NN_WC setup in the middle to lower troposphere (the NN_SC setup performs consistently worse). The picture is more nuanced for radiosonde observations, where the ANN 4D-Vars appear to perform significantly better than current weak and strong-constraint 4D-Var 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 initialized by a long-running preoperational WC-4DVar and thus are likely to be suboptimal 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 nonbias corrected observing systems (radiosondes and 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 (Figures 10c and 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 behavior 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 WC-4DVar. 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 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). Ten-day forecasts are initialized using the analysis from strong-constraint 4D-Var, WC-4DVar, 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 WC-4DVar 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 root-mean-square error (RMSE) after 72 h between forecasts initialized by NN_SC and strong-constraint 4D-Var (a) and by NN_WC and strong-constraint 4D-Var (b) and by WC-4DVar and strong-constraint 4D-Var (c). The improvements obtained by WC-4DVar 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 WC-4DVar 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 4D-Var 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 WC-4DVar, 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 localized 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 4D-Var methodology (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, et al., 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 (e.g., stochastically perturbed parametrization tendency [SPPT], stochastic kinetic energy backscatter [SKEB], and others; see Leutbecher et al., 2017, for a discussion) 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 underdispersive (Bonavita et al., 2012; Bowler, 2017; Houtekamer & Zhang, 2016). These model error parameterizations 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 recentering 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; that is, we can estimate the first moment of its distribution through statistical estimation techniques. WC-4DVar is an online example of a statistical estimation technique for dealing with the systematic errors of the model. The ML models described in this paper are examples of off-line statistical models aiming to achieve similar goals. A variety of hybrid configurations with a combination of online 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 minimization (NN_WC) is an initial, proof-of-concept attempt. Similarly to the stochastic model error parameterizations, 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. Our experience with the IFS model has been that important aspects of the model biases have a significant resolution dependence, which forces experimentation to be conducted at the target resolution. In this context it will be interesting to investigate in what measure transfer learning techniques commonly used in the ML/DL domain can ameliorate this issue in our application. 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 Crawford et al. (2020) and Bowler (2017), following earlier work by Piccolo and Cullen (2015). With respect to these later works, our application differs in some important aspects: (1) It only affects the model evolution in a deterministic sense, without adding random perturbations, and its application is limited to the data assimilation cycle; (2) the trained ANN provides a computationally compact representation of the model error information present in the available archives of analysis increments; (3) the use of an ANN gives to our estimates a nonlinear, flow-dependent capability that would be difficult to obtain from purely climatological regressions. With regards to the first point, we have not yet tested the system in an ensemble configuration, but this is firmly in our plans. Our expectation is that the use of the ANN will also have a positive effect on the reliability of the ensemble forecast through the reduction of the ensemble mean RMSE. The third property (flow dependency of the model error tendencies) is important first because the flow-dependent component adds predictive power to that coming from climatological predictors only (section 2) and also because it opens up 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, because the ANN is able to respond to changes in the atmospheric state and because the accuracy and reliability of ensemble forecast predictions at long forecast ranges are more 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 nonlinear 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 and 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, and sea surface temperatures). 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 briefly discuss them in the following subsection.

5.2. Statistical Regression and the Curse of Dimensionality

As a form of statistical learning, ML 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 data set) necessary to find an analog of the current state within an error distance measure $\epsilon$ has a scaling law of the form

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

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 noninteger number (i.e., a strange attractor). The exponential dependence of the training data set size on the effective attractor dimension makes a fully statistical approach to forecasting unfeasible (Van den Dool, 1994). For ML 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 (cf. Equation 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 data set. This limits the modes of variability allowed in the regression and allows to train the ML model on a relatively small data set and achieve good generalization 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 ML algorithm is allowed to search for solutions (this is called the hypothesis space in ML literature). This is where expert knowledge of the problem at hand becomes valuable as there is no ML 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. The approach 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; for example, use CNNs 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 data set.

6. Conclusions

ML and DL 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 filling 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 WC-4DVar framework (Laloyaux, Bonavita, Dahoui, et al., 2020; Laloyaux, Bonavita, Chrust, et al., 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 10 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 NWP output postprocessing and nowcasting (Gagne et al., 2017; McGovern 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 data sets, 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 paper will play a major role in a variety of data assimilation and forecasting applications.

Data Availability Statement

The input and output data of the experiments described in the paper are freely available for research purposes from ECMWF and can be requested following the procedures described online (at https://www.ecmwf.int/en/forecasts/datasets).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). Tensorflow: A system for large-scale machine learning. In 12th Usenix Symposium on Operating Systems Design and Implementation (osdi 16) (pp. 265–283). 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, 376–399. https://doi.org/10.1029/2018MS001472

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), 1916–1926.

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, 376–399. https://doi.org/10.1029/2018MS001472

Bragard, J., Carrassi, A., Molter, M., & 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. New-York: Manning Publication.

Crawford, W., Frolow, S., McIay, J., Reynolds, C. A., Barton, N., Ruston, B., & Bishop, C. H. (2020). Using analysis corrections to address model error in atmospheric forecasts. Monthly Weather Review, 148(9), 3729–3745.
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. I., McGovern, A., Haupt, S. E., Sobash, R. A., Williams, J. K., & Xue, M. (2017). Storm-based probabilistic hail forecasting with machine learning applied to convection-allowing ensembles. Weather and Forecasting, 32(5), 1819–1840.

Geer, A. J. (2020). Learning Earth system models from observations: Machine learning or data assimilation? Philosophical Transactions of the Royal Society, Submitted. https://doi.org/10.21957/7bj2811r

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

Ham, Y.-G., Kim, J.-H., & Luo, J.-J. (2019). Deep learning for multi-year ENSO forecasts. Nature, 573, 568–572.

Hamrud, M., Bonavita, M., & Isaksen, L. (2015). EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. Monthly Weather Review, 143(12), 4847–4864.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Muoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049.

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 (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). Adam: A method for stochastic optimization. International Conference on Learning Representations.

Laloyaux, P., Bonavita, M., Chrust, M., & Gürol, S. (2020). Exploring the potential and limitations of weak-constraint 4D-Var. Quarterly Journal of the Royal Meteorological Society, 1–16. https://doi.org/10.1002/qj.3891

Laloyaux, P., Bonavita, M., Dahoui, M., Farnan, J., Healy, S., Hlm, E., & Lang, S. T. K. (2020). Towards an unbiased stratospheric analysis. Quarterly Journal of the Royal Meteorological Society, 146(730), 2392–2409.

Leutbecher, M., Lock, S.-J., Ollinaho, P., Lang, S. T. K., Balsamo, G., Bechtold, P., et al. (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. I., Haupt, S. E., Karstens, C. D., Lagerquist, R., et al. (2017). 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 interdiscipliary journal of nonlinear science, 28(4), 041101.

Piccolo, C., & Cullen, M. (2015). 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). Neural networks for postprocessing ensemble weather forecasts. Monthly Weather Review, 146(11), 3885–3900.

Rodwell, M., & Jung, T. (2008). Understanding the local and global impacts of model physics changes: An aerosol example. Quarterly Journal of the Royal Meteorological Society, 134, 1479–1497.

Shepherd, T. G., Polichtchouk, I., Hogan, R., & Simmons, A. J. (2013). Report on stratosphere task force (824). ECMWF.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 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, 1402–1417. https://doi.org/10.1029/2018MS001597

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.