71^{st} International Astronautical Congress (IAC), The CyberSpace Edition, 12-14^{th} October 2020.

IAC–20–A6,7,4,x57835

A Deep Learning Approach to Space Weather Proxy Forecasting for Orbital Prediction

Emma Stevenson[a,b*], Victor Rodriguez-Fernandez[a], Edmondo Minisci[c], David Camacho[a]

[a] *School of Computer Systems Engineering, Universidad Politécnica de Madrid, Alan Turing street, 28038 Madrid, Spain*, {emma.stevenson , victor.rfernandez, david.camacho}@upm.es
[b] *School of Aeronautical and Space Engineering, Universidad Politécnica de Madrid, Plaza del Cardenal Cisneros, 3, 28040 Madrid, Spain*
[c] *Department of Mechanical and Aerospace Engineering, University of Strathclyde, 75 Montrose Street, Glasgow, United Kingdom G1 1XJ*, edmondo.minisci@strath.ac.uk
[*] Corresponding author

Abstract

The effect of atmospheric drag on spacecraft dynamics is considered one of the predominant sources of uncertainty in Low Earth Orbit. These effects are characterised in part by the atmospheric density, a quantity highly correlated to space weather. Current atmosphere models typically account for this through proxy indices such as the F10.7, but with variations in solar radio flux forecasts leading to significant orbit differences over just a few days, prediction of these quantities is a limiting factor in the accurate estimation of future drag conditions, and consequently orbital prediction. This has fundamental implications both in the short term, in the day-to-day management of operational spacecraft, and in the mid-to-long term, in determining satellite orbital lifetime. In this work, a novel deep residual architecture for univariate time series forecasting, N-BEATS, is employed for the prediction of the F10.7 solar proxy on the days-ahead timescales relevant to space operations. This untailored, pure deep learning approach has recently achieved state-of-the-art performance in time series forecasting competitions, outperforming well-established statistical, as well as statistical hybrid models, across a range of domains. The approach was found to be effective in single point forecasting up to 27-days ahead, and was additionally extended to produce forecast uncertainty estimates using deep ensembles. These forecasts were then compared to a persistence baseline and two operationally available forecasts: one statistical (provided by BGS, ESA), and one multi-flux neural network (by CLS, CNES). It was found that the N-BEATS model systematically outperformed the baseline and statistical approaches, and achieved an improved or similar performance to the multi-flux neural network approach despite only learning from a single variable.

Keywords: Solar Radio Flux, Deep Learning, Time Series Forecasting, Space Weather

1. Introduction

The dynamics of space objects orbiting in Low Earth Orbit (LEO) strongly depend on the characterisation of the uncertainties on the initial state, physical properties of the objects themselves (such as mass and shape) and properties of the atmosphere, chiefly the density. These atmospheric properties are strongly influenced by both solar and geomagnetic activities, whose forecasting is therefore of paramount importance for space operations, and whose forecast uncertainties are fundamental to properly characterise the uncertainties on the orbital states of spacecraft and space debris. As such, the prediction of these quantities has fundamental implications both in the day-to-day management of operational spacecraft such as collision avoidance [1], and also in the longer term, in re-entry prediction [2].

Typical atmospheric density models, which are used to model the dynamics of space objects, capture the space weather conditions using two types of proxies, one for the solar activity and one for the geomagnetic activity. The atmospheric density is predominantly influenced by the solar activity, or the so called Extreme Ultra Violet (EUV) irradiance [3]. The solar EUV is highly energetic and is absorbed by the upper atmosphere, which is subsequently heated up and ionised (creating the ionosphere), driving a change in the atmospheric density. However, as direct measurements of the solar EUV cannot be made on ground, such models rely on correlated proxy measures such as the F10.7 radio flux, which is a measurement of the intensity of solar radio emissions with a wavelength of 10.7 cm (a frequency of 2800 MHz) [4]. This quantity has a very long time series history, with data covering many decades, and as such is still the most common solar proxy for typical atmosphere models [5].

As a consequence, there have been a number of studies that have investigated and developed empirical time series forecasting methods and services for predicting the F10.7. Of these, there have been a variety of efforts using both statistical [6, 7, 8, 9, 10], and Machine Learning (ML) [11, 12] approaches. The popularity of machine learning in the field of space weather forecasting as a whole has grown significantly in recent years [13], owing to its ability to exploit large amounts of available data and capture non-linearity. However, unlike geomagnetic proxy forecasting, the use of these techniques is not yet the universal standard in solar proxy forecasting, with many operationally available forecasts still relying on statistical techniques. Moreover, many of the considered approaches on the machine learning side focus on the application of classical approaches such as Support Vector Regression (SVR) [11], or single layer feedforward neural networks [12].

However, within the last year there have been significant advancements in the field of time series forecasting by way of deep learning. More specifically, in [14], Oreshkin et al. presented N-BEATS, a deep residual architecture for univariate time series forecasting, which, for the first time, succeeded in outperforming winning approaches of recent forecasting competitions across a range of domains, which were all previously based on either statistical or hybrid (statistical + ML) methods. Its success is due to a unique architecture that combines a deep stack of fully-connected layers, backward and forward residual links, aggregation of the partial forecasts in a hierarchical fashion, and ensembling. Being a pure deep learning approach implies that, unlike statistical approaches, there is no expert knowledge, or ad-hoc feature engineering, required on the data itself in order to train the model.

Given the promising performance of this state-of-the-art architecture across a range of domains, in this work we apply N-BEATS to the daily prediction of the F10.7 solar proxy and examine its feasibility over forecast horizons relevant to space operations, from 3 days for activities such as collision avoidance, up to 27 days for activities such as re-entry campaigns. To the best of our knowledge, this is the first time deep residual networks have been applied to the forecasting of solar proxies. Furthermore, we extend this approach with non-intrusive uncertainty quantification using deep ensembles [15, 16]. Finally, we perform a systematic comparison of the forecasts generated using this pure deep learning approach to those generated using other data-driven approaches, both statistical and ML, and show that it can produce competitive single point forecasts, whilst using less sources of data.

The main contributions of this work can be briefly summarised as follows:

- The use of a state-of-the-art deep neural network (N-BEATS) to forecast future values of the F10.7 using only its past history, with no additional variables and no requirement for domain-specific knowledge of the data.
- The use of deep ensembles to improve the accuracy of the forecasts and to provide a measure of model uncertainty alongside the single point predictions.
- A detailed systematic comparison with operationally available forecasts, which emphasises the strengths and weaknesses of this approach and paves the way for future work. To this end, the forecasts provided by our approach, along with the code to reproduce the experiments of this paper, are publicly available on a Github repository[1], to enable further research and comparisons.

The paper is structured as follows. In Section 2 we provide backgrounds on the state of the field of time series forecasting, with a particular emphasis on its use in relation to space weather activities. In Section 3, the proposed approach is described, which includes not only the explanation of the deep learning architecture employed, N-BEATS, but also the way the data is extracted and passed to the model, the training and evaluation procedures, and the estimation of the prediction intervals through an ensemble of trained models. Section 4 makes a detailed comparison of the proposed approach with respect to current operationally available forecasts, comparing both the values of the predicted data points in the future, and the uncertainty intervals. Finally, in Section 5 we discuss the results obtained and outline avenues for future research.

2. Backgrounds on Time Series Forecasting

The goal of time series forecasting is to predict the values of a set of future data points given a set of past observations. There are multiple types of forecasting, depending on different criteria:

- The *number of series to predict*. The term univariate time series forecasting refers to making predictions on one single series, regardless the number of input variables used. On the other hand, multivariate time series forecasting refers to the prediction of several related series at once.
- The *number of time steps to predict*, also known as the horizon (H). In contrast to one-step-ahead predictions, multi-horizon forecasting predicts the variables of interest at multiple future time steps, thus providing decision makers with an estimate that can be used to optimise their course of action across an entire path of predictions. The number of time steps used to create the prediction is then known as the lookback.

[1] `https://github.com/stardust-r/deep-learning-space-weather-forecasting`

- The *uncertainty estimation* provided by the forecasting model. Single-point models focus on estimating, as precisely as possible, the future point values. However, in many scenarios [17], the provision of uncertainty intervals can be useful, if not critical, for risk management, by giving decision makers an indication of likely best and worst-base values that the target can take.

Although there have been recent attempts to create a meaningful distinction between forecasting methods [18], these can be roughly classified as being either of a statistical or machine learning nature. Statistical methods make use of statistics based on historical data to predict what will happen in the future. They are normally computationally efficient as they rely on linear processes to minimise the prediction error, and require expert knowledge about the trend and the seasonality of the data to model. Traditional and popular examples of these methods include ARIMA [19] and Exponential smoothing (ETS) [20] models. On the other hand, ML methods tackle the problem of forecasting as a supervised learning (auto)regression task, where the model is trained on pairs of past/future values from different slices of the time series. They are computationally more demanding and rely, in many cases, on non-linear training algorithms. In a "pure" ML method, the main advantage is that no time series specific engineering is needed to train the model.

Among the several ML methods that can be used for time series forecasting, neural networks, and more specifically, deep neural networks, are one of the most popular alternatives in the recent literature, due to the latest breakthroughs in Deep Learning (DL) [21]. A neural network is an artificial model that emulates how the human brain works, using an abstract (or simplified) mathematical model of a neuron. It consists of a series of such neurons connected to each other with a series of weights. These weights are learned from the training data, using a learning algorithm that updates them in order to minimise the loss (or error) of the network predictions summed over all training cases. The most common type of neural network is the feed forward neural network, where the information enters into the input units, and flows in one direction through the hidden layers until it reaches the output units. Although the universal approximation theorem [22] shows that any function can be well-approximated using a feed forward neural network with just one hidden layer of non-linear neurons, in practice, deeper architectures (with more than one hidden layer) have smaller matrices, making it possible to split the derivative of the loss function into pieces, meaning that the model can be trained more quickly and will take up less memory [23]. In addition, the layered structure of a deep neural network enables the automatic extraction of features from the data at different levels of abstraction, the later layers being the most specialised for the task at hand. In the field of time series forecasting, the most common deep learning architectures are those based on Recurrent Neural Networks (RNNs), whose units contain an internal memory state which acts as a compact summary of past information [24]. In recent years, the development of attention mechanisms and the Transformer architecture [25] has also lead to improvements in temporal dependency learning, from which time series forecasting has benefited [26].

Despite all of this, the use of ML and DL methods are far from being the standard for the task of time series forecasting. For instance, in the 2018 forecasting competition M4[2], which challenges researchers to forecast time series data over multiple domains, 12 out of the 17 most accurate solutions were ensembles of classical statistical methods [27], and only six of the submissions were pure ML. Only in the last few months, when the 2020 M5 competitions[3,4] were concluded, could it finally be seen that ML methods were part of the top solutions of the leaderboard, which represents a significant step forward in the implantation of ML for time series forecasting.

2.1 Forecasting the F10.7 proxy

Given the importance of the F10.7 proxy to atmospheric density modelling, many research works have been carried out to derive and test forecasting methods and approaches. Support Vector Regression (SVR) was used for short-term forecasting of F10.7 [11], with the authors showing that "*the proposed approach can perform well by using fewer training data points than the traditional neural network*". A simple linear forecasting model for the F10.7 proxy as been proposed by Warren et al. [8]. In this paper the authors also compared the linear forecasting approach of the F10.7 to forecasting using artificial neural networks, and preliminarily concluded that "*forecasting via sophisticate artificial neural networks is not any better than a simple linear forecasting approach*". Various empirical time series prediction techniques were compared in [12]. The authors selected a multi-wavelength, non-recursive, analogue neural network, and found that "*the prediction of the 30cm flux, and to a lesser extent that of the 10.7cm flux, performs better than NOAA's present prediction of the 10.7cm flux, especially during periods of high solar activity*". A linear multi-step forecasting model based on the correlation between different forecasting steps and the characteristic of heteroscedasticity is proposed in [9]. In the same paper, a variational Bayesian procedure to optimise the model is also in-

[2] https://www.kaggle.com/yogesh94/m4-forecasting-competition-dataset

[3] https://www.kaggle.com/c/m5-forecasting-accuracy,,

[4] https://www.kaggle.com/c/m5-forecasting-uncertainty

troduced, and it is claimed that the proposed model improves the performance of multi-step F10.7 forecasting by considering correlation and heteroscedasticity. More recently, a thorough analysis of the power of statistical ARIMA models for the forecasting of this proxy was carried out in [10], proving that, as long as the order p of the ARIMA model is optimally chosen, the model is not inferior to other techniques.

To the best of our knowledge, the reception of deep neural networks to forecast the solar flux is limited, especially in terms of the F10.7 proxy. Most of the work in the intersection of solar activity and deep learning is focused on the early classification of solar flares and geomagnetic storms. As an example, Long Short Term Memory (LSTM) architectures (a subclass of RNNs) have been employed for the detection of geomagnetic storms based on the K_p index in [28], and in [29], Convolutional Neural Networks (CNNs) are trained to classify flaring and nonflaring active regions using line-of-sight magnetograms. Only in the last few months, the task of forecasting future values of the GOES X-ray flux has been studied with different deep learning architectures [30], including N-BEATS, the architecture used as a basis for this work.

3. Applying Deep Learning to F10.7 Forecasting

In this section, we present our approach to provide daily, univariate, multi-horizon forecasts with prediction intervals of the F10.7, using only past information of the proxy, with no additional input variables. It is an end-to-end deep learning approach based on the novel architecture N-BEATS. For the sake of reproducibility, the implementation of this approach is publicly available on Github[5].

3.1 Model Architecture: N-BEATS

N-BEATS (Neural Basis Expansion Analysis for interpretable Time Series forecasting) [14] is a novel deep learning architecture for single point, univariate, multi-horizon forecasting that has been gaining traction in the field since it was proved to be the first pure deep learning method that outperforms the winning approaches of recent forecasting competitions. It does not need any specific expert knowledge on the data, and is thus applicable to a wide array of target domains without any feature engineering. An open source implementation of N-BEATS[6], written with the deep learning library PyTorch, has been employed in this work.

N-BEATS belongs to the family of deep residual networks, which were introduced for computer vision tasks as a way to train very deep networks effectively [31]. More specifically, its topology is described as doubly residual stacking (see Figure 1), where each stack consists of multiple residual blocks that produce two outputs: the block's estimation of the input (or lookback) data, called backcast, and the estimation of the future values across the desired horizon, known as forecast. The backcast is substracted from the current's block input, forming a residual which then serves as input to the the next block in the stack. The output of the network is the result of a hierarchical aggregation of the forecasts across stacks, i.e., first the partial forecasts of each block are aggregated at the stack level and then at the overall network level, providing the final global output. The iterative and residual nature of this architecture aims to encourage gradual signal reconstruction and forecasting.

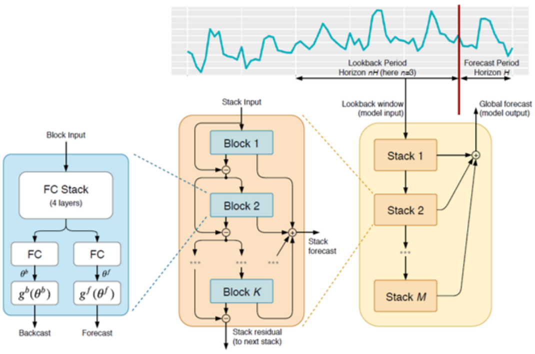


Fig. 1: N-BEATS architecture diagram, as shown in the original paper by Oreshkin et al. [14].

Internally, each basic residual block within a stack consists of a multi-layer fully connected network with non-linear (ReLu) activation functions between each layer, although there are some extensions of N-BEATS that replace this basic building block with temporal aware structures such as RNNs [32]. The fully connected network outputs two vectors of basis expansion coefficients, normally referred to as θ, which are then accepted by two learnable basis functions to generate the final backcast and forecast of the block, respectively. The parameters of the basis functions can be constrained so that only a family of functions can be learnt (e.g, low-degree polynomials or Fourier series), forcing the model to decompose the forecast into distinct human interpretable outputs such as the trend and the seasonal components of the data. On the other hand, the parameters of these functions can be left unconstrained (or generic, as it is known in the N-BEATS paper), with the aim of using no domain knowledge in the modelling.

3.2 Data and Model Inputs

The ESA Space Weather Service Network[7] maintains a database containing both past, and forecast, values of solar and geomagnetic indices which are relevant to drag calculations. This data is compiled from a number of independent providers in one place, for the

[5] See footnote 1.

[6] `https://github.com/philipperemy/n-beats`

[7] `http://swe.ssa.esa.int/`

convenience of end users and space operators, as a part of its Space Surveillance and Tracking Service.

We use this service to extract the time series of the F10.7, measured in solar flux units (sfu), which has been measured continuously since 1947 by the Ottawa, and then Penticton Radio Observatories [4]. The F10.7 is available in either *observed* values, which vary throughout the year with the Sun-Earth separation, or *adjusted* values, where the observations are adjusted to 1 AU (Astronomical Unit). In this work, we train our forecasting model using the *observed* data, which is used by typical thermosphere models [5, 12].

The data is split into training and validation subsets. However, due to the correlated nature of time series data, the typical ML strategy of random splitting, that ensures that the underlying distribution in these subsets is the same, cannot be used. We must therefore ensure that the validation set contains a full solar cycle so that it is representative of the training data. To achieve this, we use an approximate 80% to 20% splitting strategy, with the training set covering the period from 1/1/1950 to 1/1/2005, and the validation set covering 1/1/2005 to 1/1/2020, covering Solar Cycle 24, as shown in Figure 2.

As we are using deep learning, for which successive feature extraction through the layers of the architecture is implicit, this data does not require extensive pre-processing in order for the model to perform well. As such, the input data needs only to be normalised, to prevent exploding gradients and improve the numerical stability of the model, and subdivided into lookback-horizon windows. No explicit knowledge or analysis of time series features such as trend and seasonality is required in advance.

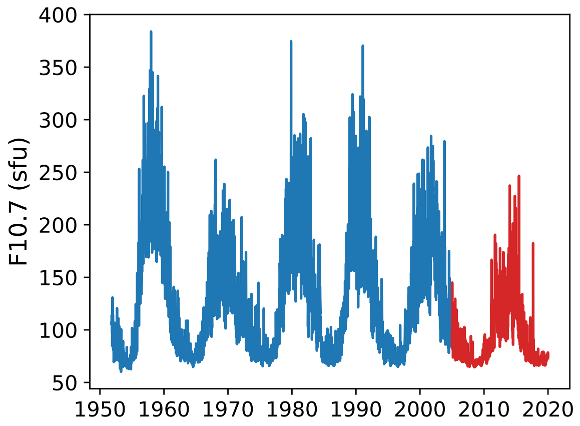


Fig. 2: Splitting of the F10.7 time series into training (blue) and validation (red) datasets.

3.3 Model Training & Evaluation

The underlying architecture of a deep learning model is defined by a series of hyperparameters which can be pre-set by the user, and are not learnt by the model during training. These parameters constrain the complexity of the model and should be optimised, or tuned, to find the optimal configuration for a specific problem such that the model does not under or overfit the training data.

In the case of the N-BEATS architecture, as can be seen in Figure 1, there are a large number of potential hyperparameters that can be tuned. However, one of the distinguishing aspects of this architecture, was that it was specifically designed to be generally applicable across a wide variety of horizons and datasets, and should perform well without the need for extensive tuning.

As such, we set architectural hyperparameters (for example the number of layers, number of stacks etc.) to those recommended by the authors [14], and focus on basis choice, lookback, and optimisation hyperparameters such as learning rate. Tuning was performed through a grid search, using the experiment-tracking tool Weights & Biases [33], and the chosen parameters are given in Table 1.

Table 1: Hyperparameter settings for the N-BEATS architecture and training procedure.

Parameter	Value
Number of Stacks	2
Basis Type	Generic
Dimension of Basis Coefficients (θ)	7,8
Share Weights in Stack	False
Number of Blocks per Stack	3
Number of Layers per Block	4
Number of Hidden Units per Layer	128
Activation Function	ReLu
Optimiser	Adam
Learning Rate	1e-4
Weight Decay	0

Notably, we found that constraining the bases to trend (polynomial) and seasonality (Fourier) components (in an interpretable approach, as described in Section 3.1) significantly hindered model performance compared to the generic approach, where the model has free reign to learn the preferred basis. By not constraining the *Basis Type*, the final model therefore does not rely on domain specific knowledge.

This analysis was performed using the Mean Squared Error (MSE), the nominal metric used to evaluate the performance of regression problems in machine learning, which is defined as follows,

$$\mathrm{MSE} = \frac{1}{H}\sum_{i=1}^{H}(\hat{y}_{T+i} - y_{T+i})^2, \qquad (1)$$

in which $\vec{\hat{y}}$ are the set of predicted future values of a time series of length T over a forecast horizon of length

H, and $\vec{y}$ are the set of true observed values over the horizon, $\vec{y} = [y_{T+1}, y_{T+2}, ..., y_{T+H}]$.

For a more robust and systematic approach to model evaluation, several additional metrics will also be considered in this work, which capture different aspects of the model performance.

Firstly, we include the Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE), which are standard scale-free metrics used more specifically in the field of time series forecasting, and used by [14] to enable performance comparison over a range of different datasets,

$$\text{MAPE} = \frac{100}{H} \sum_{i=1}^{H} \frac{|\hat{y}_{T+i} - y_{T+i}|}{|y_{T+i}|}, \tag{2}$$

$$\text{MASE} = \frac{1}{H} \sum_{i=1}^{H} \frac{|\hat{y}_{T+i} - y_{T+i}|}{\frac{1}{T+H-m} \sum_{j=m+1}^{T+H} |y_{j-m} - y_j|}. \tag{3}$$

These are linear metrics, which means that unlike the squared MSE, they do not give as much weighting or importance to larger errors, which are typically associated with higher levels of solar activity.

The MASE is equivalent to the Mean Absolute Error (MAE), scaled by the average error of a naive baseline model whose forecast is simply a previously observed value m periods in the past. If there is no prior knowledge of the seasonality of the time series, m can be set to 1 and the naive model is that of the persistence. As we do not want our analysis to depend on any preexisting domain knowledge, and as the solar flux encompasses multiple seasonalities, we consider the persistence as our baseline model here in the definition of the MASE, and later in Section 4.1.

Next, we consider the recommendations of [34], who proposed a standardised set of comparison metrics for benchmarking geomagnetic index prediction models. In this way, we hope to enable and encourage more transparent and systematic comparisons between preexisting models by future authors.

We therefore also include the Pearson linear correlation coefficient (R),

$$R = \frac{\text{cov}(\hat{y}, y)}{\sigma_{\hat{y}}, \sigma_y}, \tag{4}$$

and the Mean Error (ME), or bias,

$$\text{ME} = \frac{1}{H} \sum_{i=1}^{H} (\hat{y}_{T+i} - y_{T+i}), \tag{5}$$

which gives an indication as to whether the model, on average, overpredicts (positive bias) or underpredicts (negative bias) the observed data. We also include the MAE and Root Mean Squared Error (RMSE), the square root of Equation 1, to be consistent with the recommended guidelines, and to be comparable to other authors who may choose these metrics.

Finally, we introduce the concept of the *Relative* metric, which is an extension of a metric suggested by Yaya et al. in [12]. In the case that the model is trained separately for each horizon, using the above metrics will result in a set of performance metrics for different horizons. However, in order to obtain a single metric over all horizons, the performances must be scaled to prevent higher horizons, with higher errors, dominating the final value. We therefore define the relative metric as the average, over all horizons, of the ratio of the model performance to that of the persistence,

$$\text{Relative } X = \frac{1}{H_{\text{max}} - H_{\text{min}} + 1} \sum_{h=H_{\text{min}}}^{H_{\text{max}}} \frac{X_{\text{model},h}}{X_{\text{persistence},h}}, \tag{6}$$

where X can be any of the above metrics, and $H_{\text{min}}, H_{\text{max}}$ are the minimum and maximum horizons of interest, which in our case are 3 and 27 days respectively.

3.4 Ensemble Forecasting & Uncertainty Quantification

Ensemble forecasting is a technique that has long been used in terrestrial weather forecasting [35], and is also employed in all leading submissions in time series forecasting competitions [14], owing to its ability to not only improve accuracy, but also to improve the reliability of such forecasts by providing an inherent measure of model uncertainty [36].

This is achieved by averaging the predictions over a diverse set of models to create a single more-accurate model, with an associated uncertainty, that has several additional advantages. For example, by providing a range of possible outcomes, this approach can yield a better understanding of extreme events, such as solar storms. It can also be used to account for both uncertainty in the model inputs (aleatoric uncertainty), and the propagation of uncertainty inherent in the model itself (epistemic uncertainty, as considered in this work) in the resultant forecast uncertainty [35].

However, one of its greatest benefits when employing deep learning techniques, is its ability to improve the out-of-distribution robustness of the model [16]. Due to their large network complexity, deep learning models are particularly susceptible to overfitting, from which the model does not generalise well to new data. Regularisation techniques can be used to overcome this, and in the case of the N-BEATS architecture, ensembling was found to be more powerful than typical techniques such as drop out, or weight penalties [14].

We adopt an approach similar to that recommended in [14], building the ensemble from a set of models that have the same underlying architecture (which is chosen through hyperparameter tuning, see Section 3.3), but different higher level training parameters. For this we use three main sources of diversity: length of input window (lookback), choice of loss function (the error used internally during training, see Section 3.3), and choice of weight initialisation, as described in Table 2.

Table 2: Ensemble parameters per horizon, H. For every horizon, individual models are trained on lookback windows of different lengths, with different loss functions (as defined in Equations 1-3) and random initialisations, resulting in 90 individual models which are then aggregated to form the ensemble prediction.

Parameter	#	Values
Lookback Period	6	$[H, 2H, ..., 6H]$
Loss Function	3	MSE, MAPE, MASE
Initialisation	5	Random

In this way, the resulting ensemble can account for trends in the data over different time scales, and account for model bias and variance arising from the training procedure. This procedure is iterative and stochastic, and therefore the initial values of the weights strongly determine which local optima is found. Varying the initialisations, and the search space itself by varying the loss function, therefore improves the performance of the ensemble as a whole by averaging out weaker solutions. Injecting randomness in the initialisation is also particularly important when using machine learning techniques for time series forecasting, as the sequential nature of the data prevents the usual practice of shuffling the training dataset prior to each epoch in order to provoke changes in the gradient estimate of the optimiser.

As shown in Table 2, for every forecast horizon, we generate 90 individual models, which are then combined using the mean as the ensemble aggregation function, to generate the final forecast. In this approach, we use the same random initialisations over all horizons, and do not use bootstrapping[8], in order to ensure that each model is trained with as large a dataset as possible. Such an approach has been shown to perform well in practice compared to traditional bagging procedures [15]. It can be seen from Figure 3, that using this ensemble approach improves the performance of the overall model, and that the number of individual models we consider (90) is sufficient.

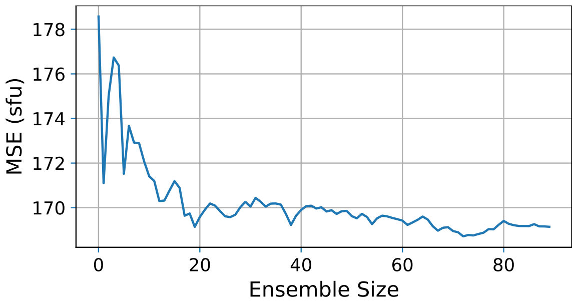


Fig. 3: Performance of N-BEATS 27-day forecast as a function of ensemble size.

An example of a 5-day N-BEATS ensemble forecast generated using the approach described in this section, which will simply be denoted as N-BEATS for the remainder of this paper, can then be seen in Figure 4. Here, we show an example of the lookback period used for the prediction (10 days, $2H$, as one of the lookbacks used in the ensemble), and the N-BEATS ensemble prediction over the forecast horizon. The 1σ uncertainty band shown here naturally arises from the distribution of forecasts over the ensemble, and can be seen to encompass the true values of the F10.7 over the horizon.

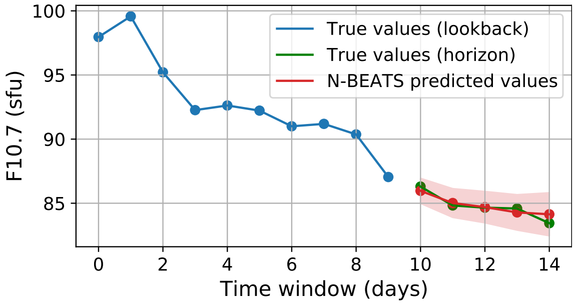


Fig. 4: An example of a 5-day forecast of the F10.7 generated by N-BEATS (true values over the forecast horizon shown in green, N-BEATS prediction in red) with an example 10 day lookback window (shown in blue). The associated 1-sigma uncertainty of the forecast generated using the ensemble approach is also shown in red.

4. Comparison with Operationally Available Forecasts

In this section, the performance of the N-BEATS ensemble approach described in Section 3, is compared to operationally available forecasts that comprise both statistical and machine learning approaches. The models themselves are described in Section 4.1, with the results of the comparisons in terms of single-point forecasting and uncertainty estimation discussed in Sections 4.2 and 4.3 respectively.

4.1 Forecast Model Descriptions

Here, we describe the F10.7 forecasts used for the comparison with N-BEATS. They are publicly avail-

[8] Bootstrapping is a resampling technique that draws samples N times uniformly with replacement from a dataset with N items

able, and updated on a daily basis.

4.1.1 Persistence (baseline)

The persistence model forecasts always the last observed value, i.e., $\hat{y}_{T+i} = y_{T+i-1} \forall i \in \{1, \ldots, H\}$. It is a simple model that performs reasonably well, and thus is a common baseline to use for comparison in multiple related works [12]. Setting $m = 1$ in the MASE metric (See Eq. 3) can be thought of as comparing a certain set of predictions against the performance of the persistence model.

4.1.2 BGS ESA

In 1993, the Geomagnetism Group of the British Geological Survey (BGS) carried out work under contract to ESA to investigate forecasting techniques for predicting solar and geomagnetic activity [6]. As part of this work, they constructed a software for the forecasting of the F10.7 proxy up to 27 days ahead. This software uses an ARIMA model [19] with 60 coefficients, which are recalculated daily to reflect changing solar and geomagnetic conditions, using the preceding two years of data.

4.1.3 CLS CNES

The Collecte Localisation Satellites (CLS), a subsidiary of the French Space Agency (CNES), provides forecasts of the F10.7 using a model developed from their research, published in [12]. Although their proposed method is similar to the approach proposed in this work in the sense that both are based on neural networks, there are two main differences:

1. Unlike the approach presented here, CLS CNES use additional input variables aside from the F10.7 to compute the forecasts. More specifically, multiple wavelengths (8.2cm, 10.7cm, 15cm and 30cm) of the solar radio flux are included.

2. The architecture employed cannot be considered as a deep neural network, since it only has one hidden layer. This layer then relies on a logistic activation function, while N-BEATS relies on ReLu. Additionally, the architecture is based on a feedforward approach, which differs to the deep residual approach used by N-BEATS.

4.2 Comparison of Single Point Forecasting

In this section, we present a comparison of the model performances in terms of single point forecasting. The analysis is comprised of two subsections.

First, we consider forecasts provided by the ESA space weather service network[9]. These forecasts, which include BGS, are only available since late 2016, and therefore this analysis is performed on a reduced validation set covering the period 1/1/2017 to 1/1/2020.

The second section then contains an extended comparison, covering a full solar cycle, between N-BEATS and CLS, whose complete forecast archive is publicly available at [10]. This covers the full validation set shown in Figure 2, from 1/1/2005 to 1/01/2020.

4.2.1 2017-2020 Validation Period

This analysis covers a period of relatively low solar activity, as seen in Figure 2, but over which we can compare N-BEATS to all the models described in Section 4.1.

In Figure 5, we show the evolution of different performance metrics, defined in Section 3.3, for these models. Figures 5a and 5b are error metrics, and so better performing models have lower errors, and as expected, the errors increase with horizon, as we forecast further forward in time. The opposite is true of the Pearson correlation coefficient in Figure 5c, where higher values are desirable, with a value of 1 indicating a perfect linear correlation between the observed and predicted values of the models.

It can be seen that the N-BEATS ensemble model gives consistently good results up to a forecast horizon of 27 days, outperforming the baseline persistence model, and the statistical BGS ESA approach in all of the displayed metrics.

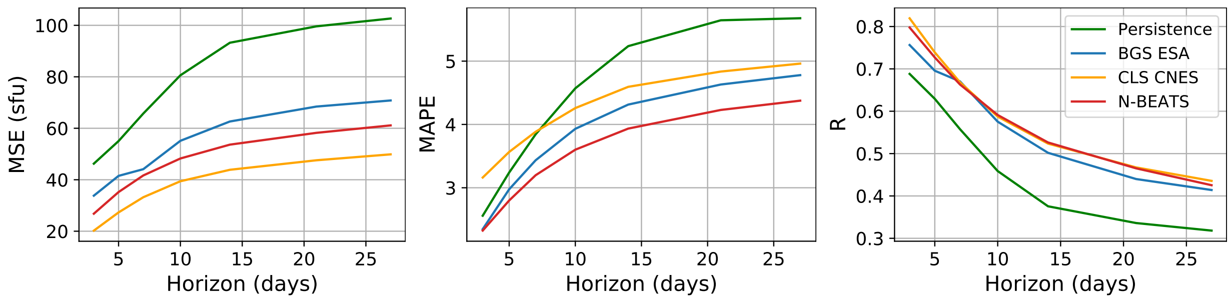
It can also be seen that the N-BEATS approach is the best performing model in terms of MAPE (Figure 5b), also outperforming the CLS CNES model. However, this behaviour is not consistent across the metrics, and it significantly underperforms, when compared to CLS, in MSE (Figure 5a). One possible explanation for this, is that the CLS model used a variation of the MSE as their loss function during training [12]. In this way, the model learns to minimise this specific metric and, as a result, forecasts generated with this model may have a bias towards it. On the other hand, we use both the MSE and MAPE as loss functions during the ensemble approach, which works to minimise bias in the final model.

This leaves the correlation coefficient, R, as the only independent comparison metric that was not used during training for either approach and, as can be seen in Figure 5c, both N-BEATS and CLS have a very similar performance.

To infer the overall best performing model, we use relative metrics, as defined in Equation 6, to obtain a set of single performance metrics which are averaged over all forecast horizons. As described in Section 3.3, these are scaled against the persistence at each horizon before they are averaged to ensure that the final metrics are not weighted too heavily towards larger, more error prone horizons, and therefore a relative metric value of

[9] http://swe.ssa.esa.int/

[10] https://spaceweather.cls.fr/



(a) Mean Squared Error (MSE) (b) Mean Absolute Percentage Error (MAPE) (c) Pearson Correlation Coefficient (R)

Fig. 5: Evolution of performance metrics with forecast horizon for 2017-2020.

1 means that the model exhibits the same performance as the persistence baseline. The model performances, in relative metrics, are given in Table 3.

Table 3: Relative metric comparison of N-BEATS with operationally available forecasts for 2017-2020. Metrics are scaled against the persistence baseline, and averaged over forecast horizons. Lower error metrics and higher correlation metrics are preferred, with a value of 1 exhibiting the same performance as the persistence baseline. The best performing values in each metric are highlighted in bold.

Model	Relative Metric				
	MSE	RMSE	MAPE	MAE	R
BGS ESA	0.698	0.836	0.867	0.875	1.230
CLS CNES	**0.480**	**0.693**	0.984	0.971	**1.285**
N-BEATS	0.601	0.775	**0.809**	**0.820**	1.273

Again, it can be concluded that N-BEATS consistently outperforms the persistence, as its relative error metrics are below, and correlation metric is above 1 respectively, but more significantly, that it systematically outperforms the BGS statistical approach.

The same conclusions can also be quantified as were previously discussed when comparing N-BEATS to CLS. It can be seen that CLS outperforms N-BEATS by at least 11% in MSE-related squared metrics, but underperforms by at least 18% compared to N-BEATS in MAPE-related linear metrics. However, the relative performance is much closer in R, where CLS outperforms N-BEATS by less than 1%.

As such, it is difficult to definitively conclude whether N-BEATS or CLS is the preferred model over this time period. It can, however, be inferred that the machine learning approaches are more powerful predictors of the F10.7 than the statistical models.

4.2.2 2005-2020 Validation Period

The same analysis was then performed over the full validation set, with the relative metrics for the available models shown in Table 4. It can be seen that, over a full solar cycle, the performance of N-BEATS is significantly closer to that of CLS in MSE, and now also fractionally exceeds it in R. This similarity in performance supports the previous analysis over the restricted validation set, and is an overall encouraging result, given that the CLS approach uses 4 different flux wavelengths during training, whereas N-BEATS learns only from a single variable.

Table 4: Relative metric comparison of N-BEATS with operationally available forecasts for 2005-2020. Metrics are scaled against the persistence baseline, and averaged over forecast horizons. The best performing values in each metric are highlighted in bold.

Model	Relative Metric				
	MSE	RMSE	MAPE	MAE	R
CLS CNES	**0.338**	**0.580**	0.837	0.804	1.136
N-BEATS	0.347	0.588	**0.772**	**0.768**	**1.137**

To investigate the strengths and deficiencies of the models over the course of the solar cycle, during different levels of solar activity, we consider the breakdown of MSE and relative MSE over the validation set. These are shown in Figure 6, alongside the observed F10.7 data during this period, to illustrate the high level of correlation between the model error and the level of solar activity itself.

From this figure, we can draw three main conclusions,

1. N-BEATS significantly outperforms CLS during low periods of solar activity.

2. The performance of the two models is fairly comparable throughout the other periods of the solar cycle, with CLS showing a slight tendency to perform better during increasing activity, and N-BEATS showing a slight tendency to perform better during decreasing activity.

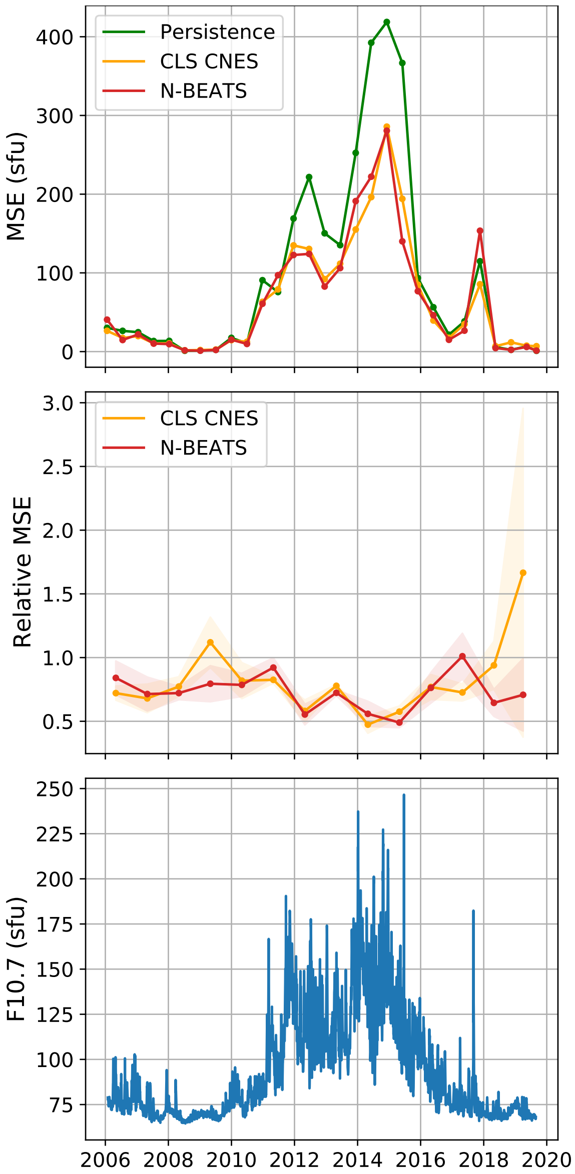


Fig. 6: Breakdown of model performances over the validation period to illustrate correlation with solar activity. *Top*: 180-day average of model MSE for a forecast horizon of 5 days. *Middle*: 365-day mean and 1σ standard deviation of the relative MSE. *Bottom*: Observed F10.7 over the validation period.

3. The models perform poorly in predicting the peak event in late 2017.

To understand these first two points, we consider the bias of the models over four year-long periods which are characteristic of low, increasing, high and decreasing levels of solar activity. These are listed in Table 5, and were chosen, where possible, to overlap with those used for a similar analysis in [12].

Firstly, it can be seen from Figure 7, that CLS is more biased than N-BEATS for low solar activity, which explains its poor relative performance during these periods. This bias is reduced during the periods of higher activity which can again be explained by its sole use of MSE-like squared metrics, which are weighted more heavily towards higher errors and therefore higher activities.

Table 5: Example periods of different levels of solar activity chosen for comparison of bias, ME, between models.

Solar Activity Level	Date Range	
Low	17/01/2008	16/01/2009
Increasing	01/05/2011	30/04/2012
High	30/09/2013	29/09/2014
Decreasing	31/08/2016	30/08/2017

N-BEATS, on the other hand, has a tendency to have better performance at lower and decreasing levels of activity. For both increasing and decreasing activity, N-BEATS has a near-zero bias up to a horizon of 5-days before it begins to underpredict, but this deviation is more pronounced during the high period of solar activity.

This systematic underprediction of high fluxes with increasing forecast horizon by N-BEATS can be better seen in Figure 8. For low horizons, the R value between the observed and predicted fluxes is close to 1. However, as the forecast horizon increases, the model tends to underestimate high values of the F10.7, resulting in a negative bias.

This damping of high activity by the model then also leads us to the third point, the peak event in 2017. This corresponds to an intense storm period that occurred during September 2017 which produced the largest flares during Solar Cycle 24 [37]. This included 4 X-class flares, the highest flare class, which have the ability to disturb satellite trajectories, and are therefore important to capture for orbital prediction [38].

It can be seen from Figure 6, that N-BEATS appears to significantly underestimate its predictions during this event. This is a result of the overconfidence of deep neural networks with out-of-distribution data, by which rare events can be erroneously predicted as in-distribution values with high confidence [15]. However, the apparent better performance by the other two models is not necessarily reliable. The persistence, for example, only performs better because it is not affected over short horizons by the problem of out-of-distribution data. As the storm period spanned several days, predicting the last known value will offer reasonable results over this period. For CLS, the error evaluation is performed against it's own provided archive of observations, which undergoes an anomaly screening. The period of observations covering this event has been tagged as “flare corrected”, and therefore may not provide a direct comparison to the N-BEATS forecast.

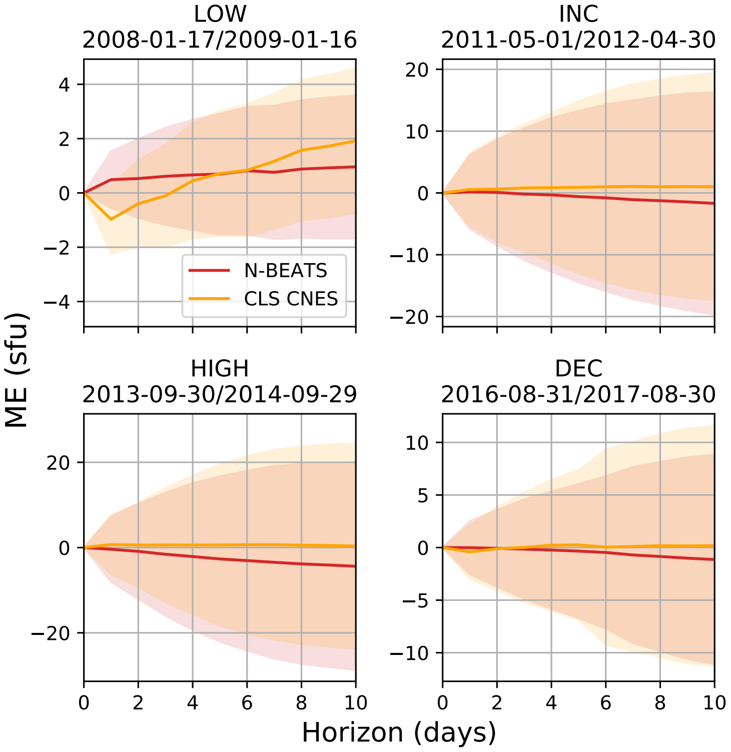


Fig. 7: Mean error (or bias) (predicted - observed) and 1σ standard deviation of error for N-BEATS and CLS forecasts over a 10-day forecast horizon during four periods of different solar activity (low, increasing, high, decreasing). Positive and negative values of the bias represent a tendency of the model to over and under-predict respectively.

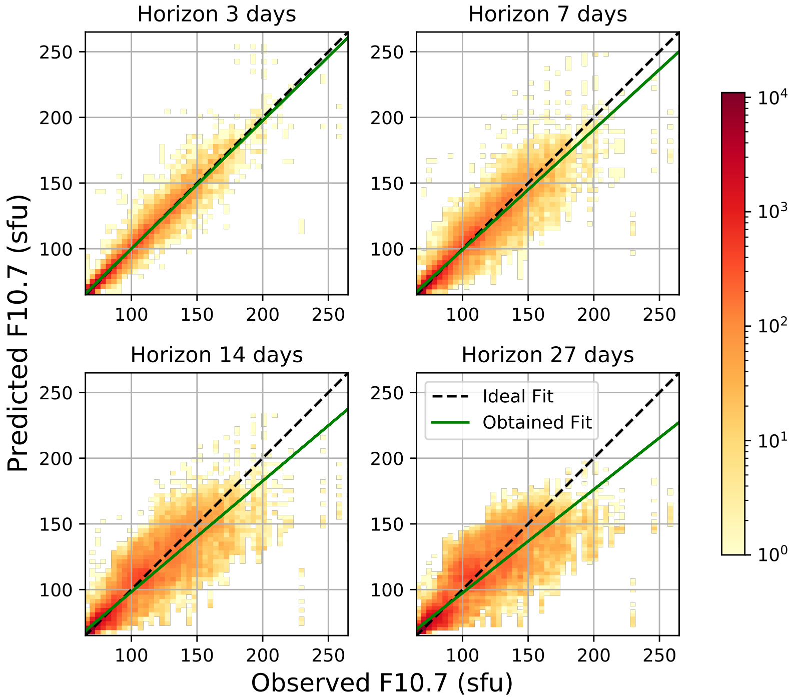


Fig. 8: Occurrence map of N-BEATS predicted values of the F10.7 with observed values over different forecast horizons for 2005-2020. The black dashed line represents the ideal $R = 1$ correlation, and the green line the fit obtained through linear regression.

4.3 Comparison of Uncertainty Estimation

The forecast uncertainty of the N-BEATS model, as described in Section 3.4, is obtained as the variance over the ensemble of individual model runs. The yearly-averaged level of this 1σ uncertainty is shown in Figure 9, in which it can be seen that the uncertainty estimation is highly correlated with the level of solar activity, and also increases with forecast horizon. From this behaviour, also observed for the single-point forecast error (Figure 6), we can therefore conclude that this approach correctly characterises the shape of the forecast uncertainty.

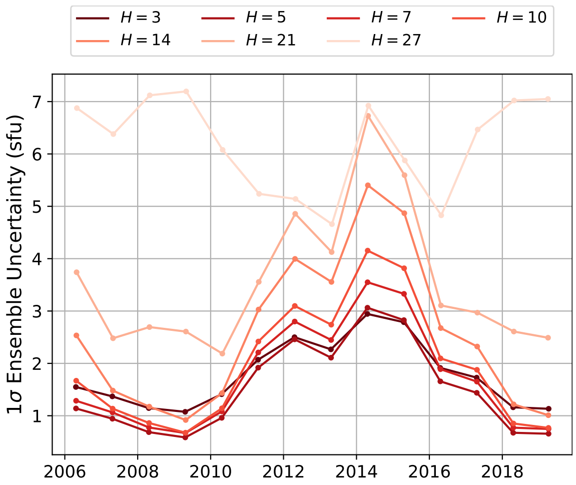


Fig. 9: Yearly-averaged forecast uncertainty of N-BEATS over various forecast horizons, H (days). Forecast uncertainty taken to be the 1σ standard deviation over the ensemble.

It can, however, be seen that although well-correlated to the periods of high solar activity, the 27-day, and to a lesser extent the 21-day forecast horizons, deviate from the solar activity profile during lower activity. This suggests that a significant number of the individual models that are aggregated to form the ensemble model for these horizons are over-predicting the flux, which affects the performance more notably during these periods. We believe that this behaviour can be attributed to individual models with very high lookbacks, as these are a function of the horizon, for example $5, 6H$. Such a broad lookback window is evidently detrimental for the modelling, since it causes too large a difference in the input data used by each member of the ensemble, which makes the uncertainty less accurate. In fact, the reason why the initial uncertainties (covering 2006-2008) for $H = 27$ are so high during a period of low solar activity is likely due to the model using input data from past periods of higher activity (end of 2005, see Figure 2). Therefore, the 27-day horizon model may not be reliable with this lookback ensemble strategy, so it will be excluded from the following analysis.

Having verified the shape of the estimated uncertainty, we can now compare its relative magnitude to that of CLS, whose forecast service also provides a measure of uncertainty. Unlike our approach, this is assigned using a linear fit between past model RMSE and solar flux [12].

A comparison of the forecast uncertainties of the two models is shown in Figure 10 for a 14-day horizon, though it should be emphasised that the same relative behaviour is true of all horizons. It can be seen that both follow the same distinct trend, but that the N-BEATS estimation is much narrower than that of CLS.

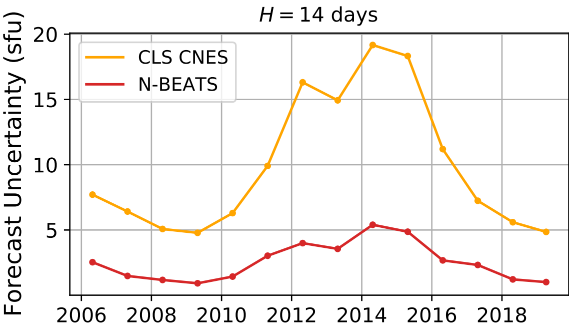


Fig. 10: Comparison of yearly-averaged forecast uncertainties for a 14-day forecast horizon. N-BEATS uncertainty is the 1σ standard deviation over the ensemble; CLS uncertainty is a RMSE obtained by a linear fit of past model error with solar activity [12].

In and of itself, this observation cannot be used to form conclusions about the quality of uncertainty estimation in either approach. However, as both models have similar single-point predictive capabilities, it does suggest that one or other is under or over estimating the uncertainty.

To quantify this, we consider a simple metric that measures whether the observed, or true, values over the horizon are contained within the interval of the predicted values ± the estimated uncertainty. Averaging over the horizons, we find that this is true of the CLS forecasts 99.23% of the time, over 3σ, whereas it is only true 76.04% of the time for N-BEATS (in which the 27-day estimates were excluded as they would falsely inflate this metric).

It can therefore be concluded that the N-BEATS estimation is too narrow to accurately capture the forecast uncertainty, which is better characterised by CLS.

However, this result is not unexpected, as MSE-based deep ensembles have been found to routinely yield overoptimistic uncertainty estimates in the field of deep learning [15]. These findings should therefore not be used to dismiss the versatile nature of this preliminary approach, which was able to well characterise the shape of the uncertainty, and, unlike CLS, is a direct output of the deep learning model. In fact, suggested modifications for improved performance are not significant, and simply involve updating the loss functions used during training to those that are able to also capture the quality of the predictive uncertainty of the model - the variance as well as the mean [15]. This will be further investigated in future work.

5. Conclusions and Future Work

This paper presents the use of the novel N-BEATS deep residual network for the daily prediction of the F10.7 solar proxy. This pure deep learning approach, which has provided a significant advancement in the field of time series forecasting within the last year, was found to be effective in this task up to a forecast horizon of 27-days, without the need for any specific expert knowledge of the data or feature engineering.

Forecasts generated using this deep, univariate approach were compared to a persistence baseline and two operationally available forecasts: BGS (a statistical approach) and CLS (a shallow neural network approach based on 4 flux wavelengths). It was found that the N-BEATS model systematically outperformed the baseline and statistical approaches, and achieved an improved or similar performance to CLS in all evaluation metrics, despite only learning from a single variable. Therefore not only was N-BEATS found to be a more powerful architecture for predicting the F10.7, requiring less data to achieve the same level of performance, but this approach also has fewer sources of uncertainty.

To capture the uncertainty in the modelling, the N-BEATS model was extended in this work to provide forecast uncertainties using deep ensembles. This was shown to be a promising preliminary approach, by validating the uncertainty distribution against that of CLS, however was found to produce overoptimistic estimates at this stage.

In future work, we would therefore like to augment our model with improved uncertainty estimators, and expand the approach to produce probabilistic forecasts. We would also like to further improve the accuracy of the model by including auxiliary variables, such as additional flux wavelengths, which may also aid in correcting the tendency of the model to underpredict the flux at high solar activity. In a further step, this approach could then be used to extend the forecasting to these other variables, for example the F30, in a multivariate approach.

Acknowledgments

This research is supported by the EU H2020 MSCA ITN Stardust-R, grant agreement 813644.

The authors would also like to acknowledge CLS, BGS and NOAA, as well as the ESA Space Situational Awareness Programme and Dominion Radio Astrophysical Observatory (Penticton, Canada) for providing the data necessary to carry out this work.

References

[1] Charles D. Bussy-Virat, Aaron J. Ridley, and Joel W. Getchius. Effects of Uncertainties in the Atmospheric Density on the Probability of Collision Between Space Objects. *Space Weather*, 16(5):519–537, May 2018.

[2] B. Bastida Virgili, S. Lemmens, E. Stevenson, and B. Reihs. Statistical comparison of ISO recommended thermosphere models and space weather proxy forecasting on re-entry predictions. In *Proceedings of the International Astronautical Congress, IAC*, 2017.

[3] A. Vourlidas and S. Bruinsma. EUV Irradiance Inputs to Thermospheric Density Models: Open Issues and Path Forward. *Space Weather*, 16(1):5–15, 2018.

[4] K. F. Tapping. The 10.7 cm solar radio flux ($F_{10.7}$). *Space Weather*, 11(7):394–406, July 2013.

[5] David A. Vallado and David Finkleman. A Critical Assessment of Satellite Drag and Atmospheric Density Modeling. *Acta Astronautica*, 95:141–165, February 2014.

[6] R. Mugellesi-dow, D. J. Kerridge, T. D. G. Clark, and A. W. P. Thompson. SOLMAG: an operational system for prediction of solar and geomagnetic activity indices. In *Proceedings of the First European Conference on Space Debris*, 1993.

[7] W. Kent Tobiska, S. Dave Bouwer, and Bruce R. Bowman. The development of new solar indices for use in thermospheric density modeling. *Journal of Atmospheric and Solar-Terrestrial Physics*, 70(5):803–819, March 2008.

[8] Harry P. Warren, John T. Emmert, and Nicholas A. Crump. Linear forecasting of the F10.7 proxy for solar activity. *Space Weather*, 15(8):1039–1051, 2017.

[9] Z. Wang, Q. Hu, Q. Zhong, and Y. Wang. Linear multistep $f10.7$ forecasting based on task correlation and heteroscedasticity. *Advancin Earth and Space Science*, 5(12):863–874, 2018.

[10] Zhanle Du. Forecasting the daily 10.7 cm solar radio flux using an autoregressive model. *Solar Physics*, 295(9):1–23, 2020.

[11] C. Huang, D.D Liu, and J.S. Wang. Forecast daily indices of solar activity, $f10.7$, using support vector regression method. *Research in Astronomy and Astrophysics*, 9(6):694–702, 2009.

[12] Philippe Yaya, Louis Hecker, Thierry Dudok de Wit, Clémence Le Fèvre, and Sean Bruinsma. Solar radio proxies for improved satellite orbit prediction. *Journal of Space Weather and Space Climate*, 7:A35, 2017.

[13] E. Camporeale. The Challenge of Machine Learning in Space Weather: Nowcasting and Forecasting. *Space Weather*, 17(8):1166–1207, August 2019.

[14] Boris N. Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. N-BEATS: neural basis expansion analysis for interpretable time series forecasting. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[15] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6402–6413. Curran Associates, Inc., 2017.

[16] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[17] Pierre Pinson. *Estimation of the uncertainty in wind power forecasting*. PhD thesis, École Nationale Supérieure des Mines de Paris, 2006.

[18] Tim Januschowski, Jan Gasthaus, Yuyang Wang, David Salinas, Valentin Flunkert, Michael Bohlke-Schneider, and Laurent Callot. Criteria for classifying forecasting methods. *International Journal of Forecasting*, 36(1):167–177, January 2020.

[19] GEORGE EP Box, Gwilym M Jenkins, and G Reinsel. Time series analysis: forecasting and control holden-day san francisco. *BoxTime Series Analysis: Forecasting and Control Holden Day1970*, 1970.

[20] Rob J Hyndman, Anne B Koehler, Ralph D Snyder, and Simone Grose. A state space framework for automatic forecasting using exponential smoothing methods. *International Journal of forecasting*, 18(3):439–454, 2002.

[21] Xueheng Qiu, Le Zhang, Ye Ren, Ponnuthurai N Suganthan, and Gehan Amaratunga. Ensemble deep learning for regression and time series forecasting. In *2014 IEEE symposium on computational intelligence in ensemble learning (CIEL)*, pages 1–6. IEEE, 2014.

[22] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[23] S. Gugger and J. Howard. *Deep Learning for Coders with Fastai and PyTorch: AI Applications Without a PhD*. O'Reilly Media, Incorporated, 2020.

[24] Bryan Lim and Stefan Zohren. Time series forecasting with deep learning: A survey. *arXiv preprint arXiv:2004.13408*, 2020.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[26] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. *arXiv:1912.09363 [cs, stat]*, December 2019. arXiv: 1912.09363.

[27] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4):802–808, 2018.

[28] Yao Tan, Qinghua Hu, Zhen Wang, and Qiuzhen Zhong. Geomagnetic Index Kp Forecasting With LSTM. *Space Weather*, 16(4):406–416, 2018.

[29] Shamik Bhattacharjee, Rasha Alshehhi, Dattaraj B. Dhuri, and Shravan M. Hanasoge. Supervised convolutional neural networks for classification of flaring and nonflaring active regions using line-of-sight magnetograms. *The Astrophysical Journal*, 898(2):98, Jul 2020.

[30] Sumi Dey and Olac Fuentes. Predicting solar x-ray flux using deep learning techniques. In *International Joint Conference on Neural Networks (IJCNN), IEEE World Congress on Computational Intelligence (WCCI)*, 2020.

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[32] Harry Rubin-Falcone, Ian Fox, and Jenna Wiens. Deep residual time-series forecasting: Application to blood glucose prediction. In Kerstin Bach, Razvan C. Bunescu, Cindy Marling, and Nirmalie Wiratunga, editors, *Proceedings of the 5th International Workshop on Knowledge Discovery in Healthcare Data co-located with 24th European Conference on Artificial Intelligence, KDH@ECAI 2020, Santiago de Compostela, Spain & Virtually, August 29-30, 2020*, volume 2675 of *CEUR Workshop Proceedings*, pages 105–109. CEUR-WS.org, 2020.

[33] Lukas Biewald. Experiment tracking with weights and biases, 2020. Software available from `https://www.wandb.com/`.

[34] Michael W. Liemohn, James P. McCollough, Vania K. Jordanova, Chigomezyo M. Ngwira, Steven K. Morley, Consuelo Cid, W. Kent Tobiska, Peter Wintoft, Natalia Yu Ganushkina, Daniel T. Welling, Suzy Bingham, Michael A. Balikhin, Hermann J. Opgenoorth, Miles A. Engel, Robert S. Weigel, Howard J. Singer, Dalia Buresova, Sean Bruinsma, Irina S. Zhelavskaya, Yuri Y. Shprits, and Ruggero Vasile. Model Evaluation Guidelines for Geomagnetic Index Predictions. *Space Weather*, 16(12):2079–2102, 2018.

[35] Sophie A. Murray. The Importance of Ensemble Techniques for Operational Space Weather Forecasting. *Space Weather*, 16(7):777–783, 2018.

[36] J. A. Guerra, S. A. Murray, and E. Doornbos. The Use of Ensembles in Space Weather Forecasting. *Space Weather*, 18(2), 2020.

[37] P. C. Chamberlin, T. N. Woods, L. Didkovsky, F. G. Eparvier, A. R. Jones, J. L. Machol, J. P. Mason, M. Snow, E. M. B. Thiemann, R. A. Viereck, and D. L. Woodraska. Solar Ultraviolet Irradiance Observations of the Solar Flares During the Intense September 2017 Storm Period. *Space Weather*, 16(10):1470–1487, 2018.

[38] Florent Deleflie, Kelsey Doerksen, Carine Briand, Muhammad Ali Sammuneh, and Luc Sagnières. Atmospheric Density Variations and Orbit Perturbations in Relation to Isolated Solar X-flare Events. In *European Geophysical Union General Assembly*, 2019.