Environmental Data Science (2022), 1: e2, 1–17
doi:10.1017/eds.2022.1

APPLICATION PAPER

Predicting years with extremely low gross primary production from daily weather data using Convolutional Neural Networks

Aris Marcolongo[1,2,*], Mykhailo Vladymyrov[1,3], Sebastian Lienert[2,4], Nadav Peleg[5], Sigve Haug[1] and Jakob Zscheischler[2,3,6]

[1]Mathematical Institute, University of Bern, Bern, Switzerland
[2]Climate and Environmental Physics, University of Bern, Bern, Switzerland
[3]Theodor Kocher Institute, University of Bern, Bern, Switzerland
[4]Oschger Centre for Climate Change Research, University of Bern, Bern, Switzerland
[5]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland
[6]Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
*Corresponding author. E-mail: aris.marcolongo@math.unibe.ch

Received: 15 July 2021; **Revised:** 08 October 2021; **Accepted:** 07 January 2022

Keywords: Carbon cycle; Convolutional Neural Networks; extreme events; extreme impacts

Abstract

Understanding the meteorological drivers of extreme impacts in social or environmental systems is important to better quantify current and project future climate risks. Impacts are typically an aggregated response to many different interacting drivers at various temporal scales, rendering such driver identification a challenging task. Machine learning–based approaches, such as deep neural networks, may be able to address this task but require large training datasets. Here, we explore the ability of Convolutional Neural Networks (CNNs) to predict years with extremely low gross primary production (GPP) from daily weather data in three different vegetation types. To circumvent data limitations in observations, we simulate 100,000 years of daily weather with a weather generator for three different geographical sites and subsequently simulate vegetation dynamics with a complex vegetation model. For each resulting vegetation distribution, we then train two different CNNs to classify daily weather data (temperature, precipitation, and radiation) into years with extremely low GPP and normal years. Overall, prediction accuracy is very good if the monthly or yearly GPP values are used as an intermediate training target (area under the precision-recall curve AUC ≥ 0.9). The best prediction accuracy is found in tropical forests, with temperate grasslands and boreal forests leading to comparable results. Prediction accuracy is strongly reduced when binary classification is used directly. Furthermore, using daily GPP during training does not improve the predictive power. We conclude that CNNs are able to predict extreme impacts from complex meteorological drivers if sufficient data are available.

Impact Statement

Understanding and predicting extreme climate-related impacts is crucial to constrain climate risk. This is a difficult task because of the typically multiple involved time scales and interactions between impact drivers. Here, we employ Convolutional Neural Networks (CNNs) and test their ability to predict years with extremely

This research article was awarded an Open Data badge for transparent practices. See the Data Availability Statement for details.

low carbon uptake (a proxy for vegetation mortality) from daily weather data. The employed CNNs can distinguish well between normal years and years with extremely low carbon uptake, with prediction power increasing from high to low latitudes. This highlights that deep learning can be used to learn very complex relationships between daily weather data and extreme impacts.

1. Introduction

Climatic hazards such as heavy precipitation, storms, droughts, and heatwaves often have disruptive impacts on human societies and ecosystems. Heavy precipitation can cause floods, which may lead to infrastructural damages and fatalities. Heatwaves adversely affect human health. Droughts diminish agricultural output and ecosystem productivity. Despite extensive knowledge of climate extremes and associated impacts, severe climate-related impacts often surprise us and can supersede the coping capacity of the impacted system. The risk of this happening is particularly large in systems for which there is no direct correspondence between the impact and a well-defined climatic hazard (Smith, 2011; Leonard et al., 2014; Ben-Ari et al., 2018).

Natural systems not only respond to climate extremes, but also to the continuous occurrence of weather conditions on top of longer-term climate trends. Consequently, repeated moderate weather that adversely affects a system may accumulate and ultimately pass the coping capacity of that system, resulting in a large impact (Leonard et al., 2014). An example of this was the high-impact lake flood event that took place in southern Switzerland in October 2010. Two storms preconditioned the catchment and brought the lake close to its flood level. Only during the third storm, the lake level rose above the flood threshold (Lenggenhager et al., 2019). Such temporally compounding effects (Zscheischler et al., 2020) are also relevant for ecosystems, whose sensitivity to weather conditions and climate extremes depends on the vegetation composition and varies between seasons (Frank et al., 2015). Climate extremes can have a positive or negative effect on ecosystem productivity, depending on the ecosystem and when they occur (Sippel et al., 2016; Flach et al., 2018). As an example, the extremely hot and dry conditions during the 2010 Russian heatwave led to a significant decrease in photosynthetic carbon uptake in crop-dominated ecosystems in the south, whereas the energy-limited forest ecosystems in the north responded with a significant increase in uptake (Flach et al., 2018). Similarly, a heatwave in spring can lead to higher productivity, whereas a heatwave in summer may reduce productivity (Wolf et al., 2016). Despite the relevance of climate extremes, extreme impacts may be the result of an unfortunate combination of not very extreme weather and climate conditions (Van der Wiel et al., 2020). For instance, the extreme wheat loss in 2016 in France, which resulted in wide-ranging implications for the French agricultural sector, is linked to the compounding effect of warm temperature in late autumn and abnormally wet conditions in the following spring (Ben-Ari et al., 2018).

Compound weather and climate events have been defined recently as the combination of multiple climate drivers and/or hazards that contributes to societal or environmental risk (Zscheischler et al., 2018). Climate drivers may span multiple spatial and temporal scales. Identifying which combinations of climate conditions lead to an impact is a challenging task, especially in systems where the impact is a complex function of weather conditions over many time-scales, such as agriculture or natural ecosystems (Vogel et al., 2021). In particular, compounding effects of weather conditions that lead to state-dependent extreme system response, for instance, in ecosystems, cannot be identified by solely studying climate extremes (Van der Wiel et al., 2020). Uncovering such unusual conditions in the very high-dimensional climate space requires new data-analytic approaches.

Discovering new (unknown) compounding drivers of large impacts from observations is extremely difficult due to the lack of consistently high-quality and long-term datasets on climate-related impacts or impact proxies. Impacts in this context may refer to low crop yields, heat-related mortality, low vegetation carbon uptake, fire intensity, and flood height and damages, among others. Well-calibrated process-based climate impact models (Frieler et al., 2017), such as physical hydrological models, crop models, and dynamic vegetation models are an alternative to study climate-related impacts on ecosystems and different aspects of human society.

Process-based impact models can be used to question current paradigms and search for weather conditions that lead to large impacts. A hydrological model, for example, can be used to study the relationship between different precipitation patterns and flooding (Zischg et al., 2018; Peleg et al., 2020). Impact models can also be used to study the impacts of climate extremes (Bastos et al., 2020) as well as to explore and test new hypotheses. For instance, a dynamic vegetation model has been used to investigate whether increased carbon uptake in spring can compensate for reduced carbon uptake during very hot and dry summers (Sippel et al., 2017). Due to the well-constrained boundary conditions and the possibility to create near-infinite amounts of data, process-based models are an excellent tool to develop and validate new statistical approaches. Furthermore, insights from established impact models that incorporate the most relevant processes can result in new knowledge about the real world or uncover model deficiencies. In principle, one could use impact models and run many factorial simulations by only varying one driver at a time to identify weather conditions leading to extreme impacts. However, due to the extremely high number of potential impact drivers, this approach cannot be exhaustive.

As a first step toward identifying weather features associated with large impacts, we aim to predict years with extremely low carbon uptake from daily weather data. Identifying combinations of weather patterns that lead to extremely low annual carbon uptake would help to better understand ecosystem vulnerability to adverse weather conditions, and to potentially build predictive models of vegetation mortality. In particular, we aim at the prediction of gross primary production (GPP), measuring the amount of carbon extracted by plants from the atmosphere and incorporated into their structure. The total yearly GPP, rather than the individual daily scale values, measures the impact on the ecosystem of unfavorable meteorological conditions.

The simple statistical learning approaches, such as linear regression and Random Forest (Zscheischler et al., 2016; Vogel et al., 2021) are restricted in their ability to deal with high-dimensional relationships, requiring specification of suitable input features (i.e., effectively performing a dimensionality reduction as a preprocessing step). Instead, here we apply a deep learning approach (Artificial Neural Networks, ANNs) as it was shown that deep learning models are able to capture nonlinear and multivariate relationships in high-dimensional datasets, and achieve state-of-the-art performance in various fields (LeCun et al., 2015). Since ANNs typically require large amounts of training data to achieve good performance, here we use a dynamic vegetation model to create large amounts of training data at four different sites, representative of different climates and vegetation types.

2. Data

In this work, we consider three different geographic locations, which are shown in Figure 1. These climate sites are denoted as *temperate, boreal*, and *tropical*, following their respective main climatic zone. To produce representative meteorological data for each location, we used the AWE-GEN stochastic weather generator model (Fatichi et al., 2011). The weather generator simulates hourly precipitation (PPT), air temperature (AT), and radiation (photosynthetically active radiation, PAR) so that hourly-to-seasonal dynamics, as well as seasonal- and inter-annual variability, are preserved. For each site, hourly climate data obtained from the ERA5 climate reanalysis product (Hersbach et al., 2020) were used to parameterize the model. Our next step was to simulate a stationary climate ensemble of 100,000 years that represents a pseudo-replication of the current climate for each site. For more information about the model and the simulated climate stochasticity, see Fatichi et al. (2016) and Peleg et al. (2019).

The generated meteorological data are then used to drive three single grid-cell simulations with the dynamic global vegetation model LPX-Bern v1.4 (Lienert and Joos, 2018). In addition to daily temperature, precipitation, and radiation, the vegetation model is supplied with the 2010 levels of global atmospheric CO_2 concentration (387.98 ppm) and Nitrogen deposition (Tian et al., 2018). The model represents vegetation with plant functional types (PFTs), the relative abundance of which are determined dynamically by climatic conditions and competition for light, nutrients, and water. Here, only natural vegetation is considered, internally represented by eight tree-PFTs and two grass-PFTs. The leaf phenology of deciduous trees is determined by considering the warmest and coldest month in a model

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

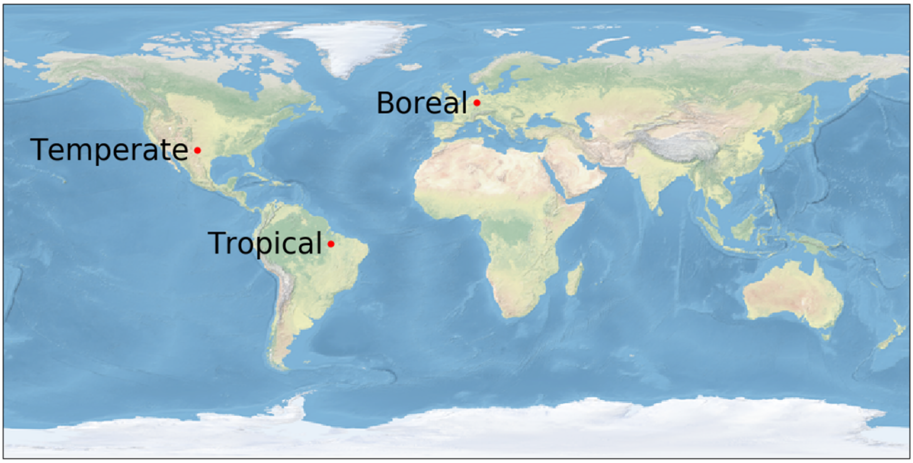

Figure 1. *Geographic locations of the three sites considered in this work. Sites are named according to their climate type.*

year. Since this can potentially lead to a dependence of GPP on the temperature in the future, we adapted the model code such that the phenology is always calculated assuming January is the coldest and July is the warmest month of the year.

The three simulations differ in their vegetation composition (Figure 2a). The temperate site is dominated almost exclusively by grasses (TrH) (we note that, despite the misleading name, in the LPX model grasses are named tropical [TrH] according to a particular photosynthetic pathway [C4] and do not necessarily grow exclusively in tropical climates Sitch et al., 2003). The boreal site is dominated by boreal trees (BoNE, BoNS, and BoBS) with few temperate trees (TeBS) and grasses (TeH), whereas the tropical site is dominated by tropical trees (TrBE and TrBR) and a few grasses (TrH). We observe that the vegetation distributions do not overlap and are representative of distinct ecosystems. The corresponding monthly GPP is reported in Figure 2b. In boreal climate, GPP has a pronounced seasonal cycle and is close to zero during colder months. Under a tropical climate, GPP is almost constant during the majority of the year but drops in the months of August and September during the dry season, during which the interannual variability is increased. The temperate site shows the highest interannual variability in GPP consistently for all months. In Figure 2c, we report the distribution of yearly GPP, showing an unstructured uni-modal distribution for all sites. In the Supplementary Information, we report daily distributions for both meteorological variables and gross primary production (Figures S1–S4).

3. Methodology

3.1. Fundamental idea

To characterize meteorological conditions leading to a low yearly GPP, we present machine learning models able to learn complex mappings between meteorological data, that is the PPT, AT, and PAR time series, considered as model inputs, and annual GPP. The developed machine learning models learn, via the provided training examples, weather patterns causing low GPP and have been devised according to the following general guidelines. First, we avoid modeling assumptions that could bias the results and their interpretation, keeping meteorological data at a daily scale as an input to our models. This way, we avoid loss of information that could result from ad hoc feature-engineering procedures neglecting day-to-day variability. Second, in order to deal with nonlinear behaviors, we exploit ANNs. Since ANNs require large number of data samples to fit their parameters, we employ simulated data that model the time evolution of weather and vegetation using a detailed process-based model (Lienert and Joos, 2018). The simulation of GPP requires a characterization of complicated biogeochemical and physical processes affecting

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

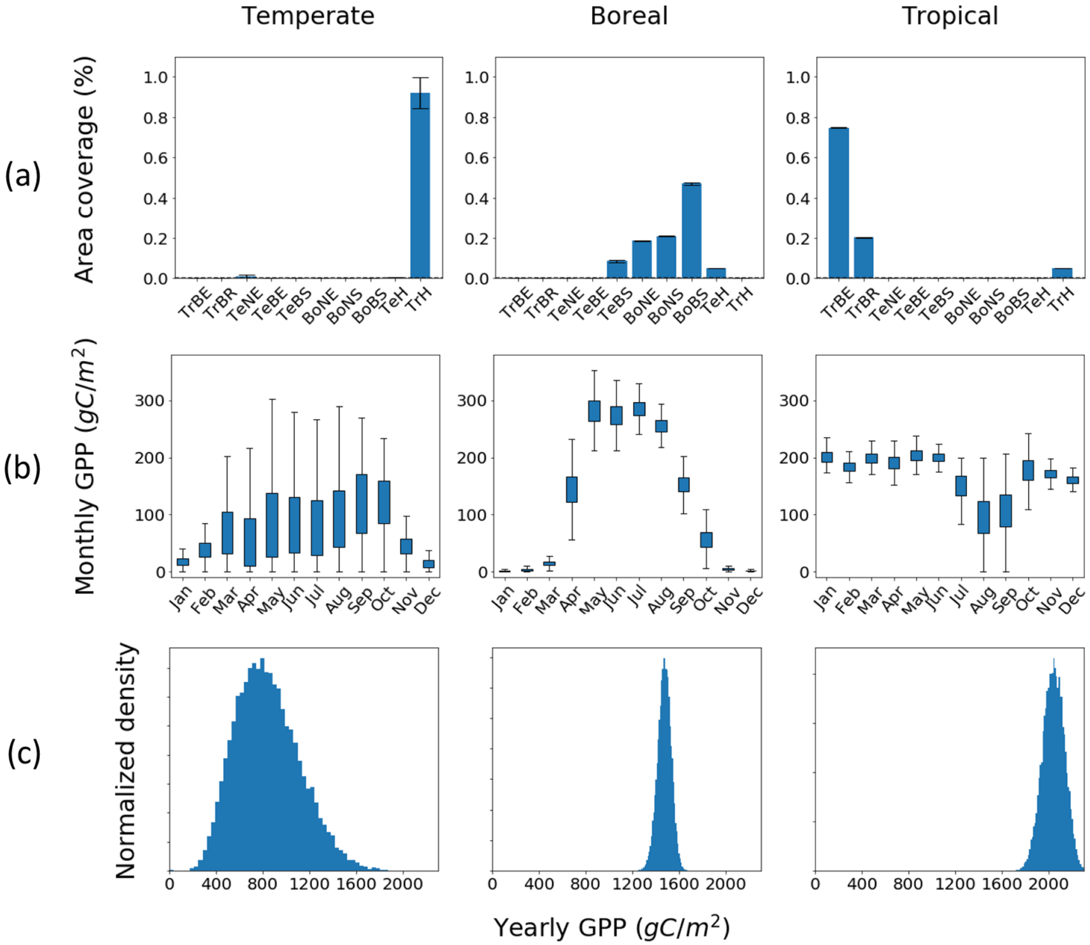

Figure 2. *(a) The surface coverage by plant functional type, as defined by the Bern-LPX vegetation model. (b) Box-plot (across years) of monthly gross primary production (GPP) values. (c) Normalized density of yearly GPP, according to which a year is defined as extreme or nonextreme.*

vegetation growth. We demonstrate that it is possible to build ANN models able to predict directly the yearly GPP from meteorological data alone. The neglected degrees of freedom, involving the exact vegetation dynamics (Lienert and Joos, 2018), will be reflected in the uncertainty of the predicted GPP.

We first construct a binary classification problem by labeling years into two classes, corresponding to extreme and normal years. Here, we define a year as extreme if the yearly GPP belongs to the lowest 10th percentile for a given site. A model can then be directly trained to classify the years into these two classes. The choice of the 10th percentile as a cutoff is somewhat arbitrary, but for our case does not affect the conclusions of the analysis. Yet, as it will be shown below, the binary approach appears to be suboptimal since it does not leverage the information available in the GPP daily time series, which in this case is used exclusively to label the years as extreme or nonextreme. A more fruitful approach consists of building intermediate regression models able to predict cumulative yearly GPP values. By fixing a cutoff on the predicted GPP values, one can trivially build a binary classifier on top of the regression model (see also Section 3.2.4). This brings the possibility to compare, using the same metrics, the performance of the classifier built on top of regression models against the classifiers trained exclusively for the classification task.

We further explore the possibility for leveraging information from the GPP time series, by building regression models which perform *intermediate predictions* of GPP aggregated at different time scales (e.g., monthly or daily values), and sum them up to obtain a yearly GPP prediction.

There are different ANN architectures that one can choose for this task. Dealing with time series, a common choice would be long short-term memory architectures (Hochreiter and Schmidhuber, 1997). Nevertheless, we believe that the complex nature of such architectures could decrease model interpretability for future analysis. Here, we focus therefore on CNNs. These architectures are often used for image analysis (Krizhevsky et al., 2012), but can be used also for one-dimensional inputs. We propose here two CNN architectures, named *CNN-M* and *CNN-D* throughout the work. The meteorological input to the models remains always the same but they differ on the time scale used for the intermediate GPP predictions (i.e., *CNN-M*—monthly and *CNN-D*—daily).

Finally, we stress that in this work, we are training distinct CNNs for each individual site.

3.2. Models

In this section, we describe the data preprocessing steps, the different model architectures, how the models were trained, and the evaluation metrics we used.

3.2.1. Data preprocessing

The PPT, AT, and PAR time series are preprocessed before being fed into *CNN-M* and *CNN-D*. AT and PAR are converted to standardized anomalies by subtracting the mean seasonal cycle. That is, for each day d and variable $X = \{\mathrm{AT}, \mathrm{PAR}\}$, $X_d \rightarrow \left(X_d - \mu_d^X\right)/\sigma_d^X$, where daily means μ_d^X and standard deviations σ_d^X are estimated over the whole time series. Daily values of PPT have distributions highly peaked at zero (no precipitation). Since we expect cumulative precipitation over several days to play an important role, we do not de-seasonalize PPT data but only rescale them with a factor fixing the daily 90th percentile $Q_{d,0.9}^{\mathrm{PPT}}$ to a value of 3 and not changing the zero value, that is $PPT_d \rightarrow 3 \times PPT_d / Q_{d,0.9}^{PPT}$. The alternative preprocessing procedure of converting PPT to anomalies, similarly to AT and PAR, leads to models with slightly worse performance. This preprocessing does not lead to loss of statistical information, since the same transformation is applied to each day independently of the year, but it ensures that the input values lie in the range compatible with standard CNNs initialization values (He et al., 2015), facilitating the training. On the output side, the GPP time series is left unchanged for *CNN-M*. For *CNN-D* the daily values of GPP are de-seasonalized on a daily bases: $\mathrm{GPP}_d \rightarrow \left(\mathrm{GPP}_d - \mu_d^{\mathrm{GPP}}\right)$.

Additionally, since the filters of the CNNs are time translational invariant, we considered inserting indexes into the network keeping track of the day of the year. This can be achieved adding additional time series to the input data. We consider the choice whether to insert these additional indexes as another meta-parameter. The day index was designed to be a time series of linearly decreasing values, where the element corresponding to the first input day has value 1, and the element corresponding to the last day of the year has value 0. We also considered two other constant input time series, with the shape of a sine and cosine, taking into account the yearly cycle. We observed that employing a combination as well as all these additional input index variables has only marginal impact on the resulting model performance for the datasets used in this work.

3.2.2. Model descriptions

3.2.2.1. CNN-M. The architecture of *CNN-M* is illustrated in Figure 3. The input consists of daily data from $T = 544$ days, corresponding to the year under consideration plus about half of the previous one. The input to the model is a 2D matrix of size $T \times N_{\mathrm{inp}}$. $N_{\mathrm{inp}} = 3$ or 6, according to whether additional indexes are used or not. The convolutional part of the *CNN-M* is composed of a series of convolution operators followed by max-pooling and produces feature maps (extracted internal representations) of dimension 34×64. These representations are then pulled together and fed into dense layers to evaluate the output. *CNN-M* first predicts some *intermediate predictions* $M_1, \ldots, M_{12}$, which corresponds to the monthly

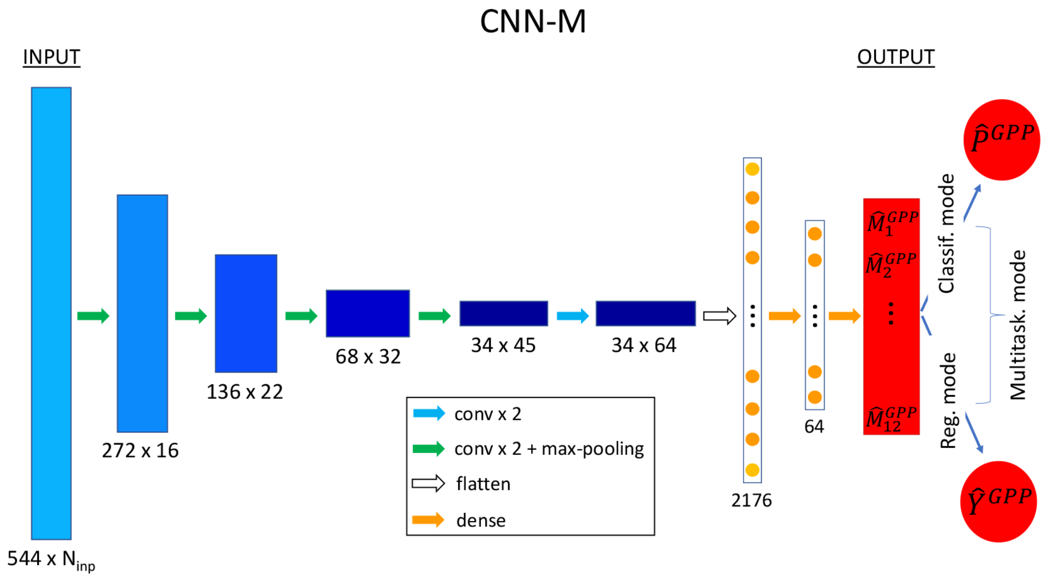

Figure 3. *CNN-M model architecture. Rectangles represent 2D or 1D tensors according to the reported dimensions. The input corresponds to meteorological data from the year considered plus around half of the previous one, for a total of 544 days. The output of the network is shown in red. Neurons $M_i, i = 1\ldots12$ correspond to monthly gross primary production (GPP) values of current year. In regression mode the monthly predictions are summed up to obtain the yearly GPP, denoted as $\hat{Y}^{GPP}$. In classification mode, a neuron predicts the probability for the year to be an extreme as well, indicated with $\hat{P}^{GPP}$. When both classification and regression modes are active CNN-M is in multitasking mode.*

amounts of GPP. The yearly GPP is evaluated as a simple sum $\sum_i M_i$. In this case, *CNN-M* is in *regression mode*. For *CNN-M*, we also explored the possibility of having a classification layer after the monthly predictions and in this case, it is in *classification mode*. By carefully weighting the loss contributions (see Section 3.2.3, Equation (5)), one can use *CNN-M* in a combined mode, to which we refer as a *multitasking mode*. We expect the multitasking approach to push the CNN toward learning features that are important for classifying years into extreme/nonextreme and neglect features that might be important to characterize the whole distribution of GPP (e.g., in an exclusive regression mode). We perform experiments to test this hypothesis.

We note that for image analysis application the number of channels in CNNs is usually multiplied by 2 after each max-pooling operation halving linear dimensions, leading to a total number of neurons that gets divided by 2. For 1D inputs, this would lead to a constant number of representations flowing through the network. We decided therefore to increment the number of channels progressively by $\sqrt{2}$, with an initial seed of 16, leading to the sequence of channels $[16, 22, 32, 45, 64]$. This way the total number of representations gets divided progressively by a factor of $\sqrt{2}$. We note that the value of $\sqrt{2}$ as a ratio between the depth of consecutive layers is arbitrary and can be chosen as a hyperparameter to tune.

3.2.2.2. CNN-D. The architecture of *CNN-D* is illustrated in Figure 4 and is a standard form of the U-Net architecture (Ronneberger et al., 2015), adapted to be used for 1D inputs. It has an encoder–decoder architecture, where the encoder part, used to build hidden representations of dimension 34 × 64, follows the same exact architecture of *CNN-M*. Differently from *CNN-M*, the resulting hidden representations go through a sequence of upsampling and convolutional operators leading to an output with the same size as the input (544 days). This way the decoder processes the high-level hidden representation found by the encoder and the prediction at each day depends on high-level features evaluated using neighboring days as well. Finally, the skip connections bring lower-level features found by the decoder closer to the output.

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

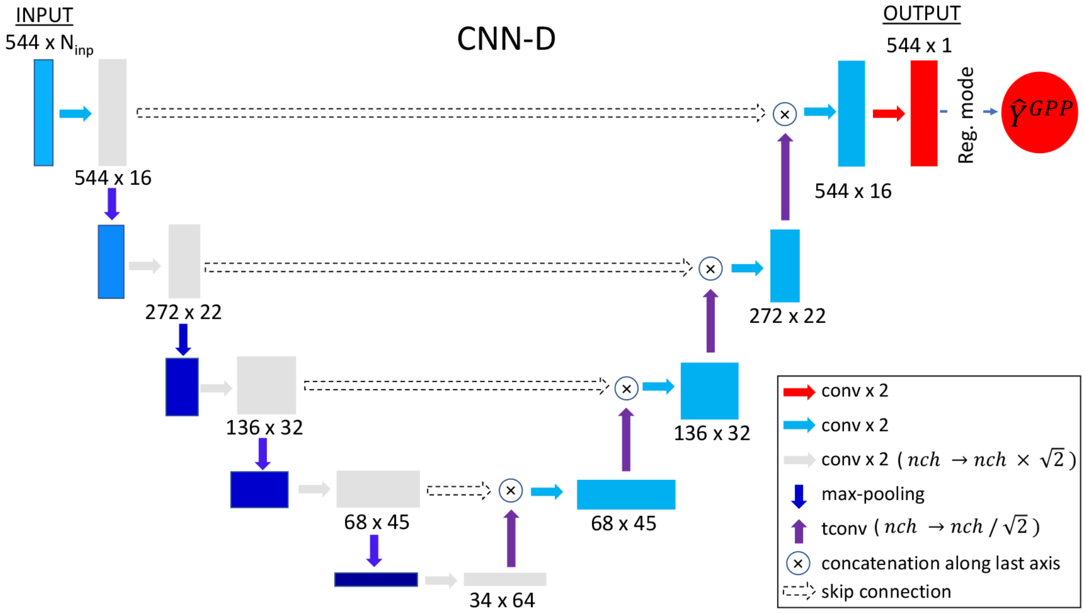

Figure 4. *CNN-D model architecture. The same notation of Figure 3 is followed. The output tensor contains daily gross primary production (GPP) values from current year plus around half of previous one (544 days),* $\hat{Y}^{GPP}$ *is the yearly GPP (sum over last 365 days).*

This not only avoids the gradient vanishing issue, but also restores the localization lost due to max-pooling operations, in order to perform predictions at the input resolution (Ronneberger et al., 2015). This way, U-net architectures are able to make daily predictions using features extracted at different time scales. From the point of view of this work, the daily outputs of *CNN-D* are considered as *intermediate predictions*, on the same ground of the monthly predictions of *CNN-M*. The last 365 days of the predictions can therefore be summed up to evaluate the yearly GPP. Only regression mode is explored for *CNN-D*.

For both CNN models the *ReLU* activation function is used (Glorot et al., 2011). L2 regularization was exploited for all weights and tuned as a meta-parameter. Last, the suffixes in the names *CNN-M* and *CNN-D* refer to the time scale of the intermediate predictions characterizing the corresponding *CNN*. Nevertheless, note that in both models by changing the meta-parameter α (see Section 3.2.3), one can choose whether these intermediate predictions should be forced to be close to the exact values or if just the final yearly output is considered.

3.2.2.3. Baseline linear model. For the baseline linear model, yearly meteorological data were first aggregated into monthly values. In particular, mean and standard deviations for each months were computed, giving a total of $12 \times 3 \times 2 = 72$ features for each year. Multilinear regression was then directly applied to predict yearly GPP. The improvement gained by including the monthly standard deviations of the meteorological data as additional features was only marginal, showing the difficulty of performing manual feature engineering for the task considered in this work.

3.2.3. Training scheme

In the following, we indicate with a hat quantities depending on the model's parameters to be optimized, for example, model outputs or loss functions. Ground truth quantities are reported without a hat. The loss function for models in regression mode, minimized through gradient descent for a batch B with N_s samples, is:

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

$$\hat{L}^{REG} = \alpha \hat{L}^{REG}_{low} + (1-\alpha)\hat{L}^{REG}_{high}, \tag{1}$$

$$\hat{L}^{REG}_{low} = \frac{1}{N_s}\sum_{s\in B}\left(\hat{Y}^{GPP}_s - Y^{GPP}_s\right)^2, \tag{2}$$

$$\hat{L}^{REG}_{high} = \frac{1}{N_s N_X}\sum_{s\in B}\sum_{i=1}^{N_X}\left(\hat{X}^{GPP}_{s,i} - X^{GPP}_{s,i}\right)^2, \quad \text{where} \tag{3}$$

$$X = M\,(\text{for } \mathit{CNN\text{-}M})\,\text{or}\; X = D\,(\text{for } \mathit{CNN\text{-}D}).$$

Here, M refers to GPP data at monthly resolution and D to GPP data at daily resolution. The functional form of the loss is equal for both *CNN-M* and *CNN-D*. $\hat{L}^{REG}_{low}$ is the loss component taking into account the low resolution part of the output (i.e., the prediction at the yearly scale) and $\hat{L}^{REG}_{high}$ takes into account the high resolution one (i.e., the intermediate predictions), corresponding to monthly or daily scale ($X = M$ or $X = D$) for *CNN-M* and *CNN-D*, respectively. Normalizations $N_M = 12$ or $N_D = 544$ are used accordingly.

The meta-parameter α weights the low and high frequency contributions to the total loss. Therefore, by choosing $\alpha = 0$ or $\alpha = 1$, it is possible to make the network fit exclusively the high-frequency or low-frequency GPP values, respectively. In particular, for $\alpha = 1$, the *intermediate predictions* are not forced to be related to any exact GPP value and only the total yearly GPP is considered from the model. The meta-parameter α can be also adjusted along the course of training.

When *CNN-M* is used in classification mode, the corresponding loss contribution is the standard cross-entropy between exact and distribution predicted by the classification output neuron:

$$\hat{L}^{CLASS} = \frac{1}{N_s}\sum_{s\in B}\sum_{c} P_s(c)\log\left(\hat{P}_s(c)\right), \tag{4}$$

where $c \in \{\text{extreme}, \text{normal}\}$ spans the two classes. If a combined classification and regression, that is a multitasking mode is used, the two losses are combined in the following way:

$$\hat{L}^{MULTI} = \beta \hat{L}^{REG} + \hat{L}^{CLASS}, \tag{5}$$

where a meta-parameter β has been introduced. When fixing $\beta = 0$, multitasking and classification models coincide. In the following for models in multitasking mode β is kept constant during training to a value selected by monitoring performance on the validation set.

The training parameters used for the optimization are reported in the Supplementary Information. We used the first 80,000 years as a training set, the following 10,000 as a validation set to tune the meta-parameters, and the last 10,000 as a test set. Results over validation and test set were always similar, indicating no overfitting to the validation set during meta-parameter tuning. More information about the training procedures is reported below and in the Supplementary Figures S9–S11.

3.2.4. Evaluation metrics

As we compare models with different outputs, choosing the right metrics is crucial for a meaningful comparison. We compare the performance of models in regression mode by plotting the distribution of yearly residuals, $\hat{Y}^{GPP} - Y^{GPP}$, in the test set. Intermediate predictions at monthly or daily scale are neglected and models are compared according to the quality of their yearly predictions. We then rescale the residuals according to the standard deviation of the exact values, $\sigma\left(Y^{GPP}\right)$. This procedure renders the residuals dimensionless and permits comparing them across different sites. From the squared sum of the rescaled residuals one can evaluate a normalized root mean squared error (NRMSE) (see also Supplementary

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

Information, Section S3). A model predicting for each year a constant equal to the standard deviation of yearly GPP has a NRMSE equal to unity. Values of NRMSE closer to zero are indicative of better performance. The same metric will be used also to evaluate the residuals in the daily GPP predicted by *CNN-D*.

In order to compare the quality of models in classification and multitasking mode, we build *precision-recall* curves using the output of the classification neuron and report the corresponding area under curve (AUC) (Sammut and Webb, 2017). These metrics are suited for binary classifiers with imbalanced classes.

Last, we compare models in regression mode against models in classification or multitasking mode using the following methodology. By fixing a cutoff on the predicted GPP values it is possible to build trivially a classification model on top of a model operating in regression mode. Varying such a cutoff one can draw a PR curve for models in regression mode as well. In practice, this can be achieved standardizing the output of the network $\hat{Y}^{GPP}$ and passing it to a sigmoid function. A PR curve can than be built with the usual routines (Pedregosa et al., 2011) if a predicted probability $\hat{P}(\text{extreme}) = \text{sigmoid}\left(\hat{Y}^{GPP}\right)$ is used. With this procedure, it is possible to compare models' performance in regression, classification, and multitasking mode on the same grounds using the PR curves obtained.

4. Results

4.1. Regression

The distribution of the rescaled residuals of the predicted yearly GPP, obtained from models in exclusive regression mode, is reported in Figure 5 and compared with the baseline linear model highlighted in green. Smaller residuals with distributions more peaked around zero imply better predictions. Our results clearly show that the prediction performance depends on the background climate and that for all sites the ANNs clearly outperform the baseline. The best NRMSE values are obtained for the tropical site, which is also the one where the linear model shows the worst results. We performed some experiments (not shown) suppressing the dry season artificially via an increased precipitation. These experiments lead in the

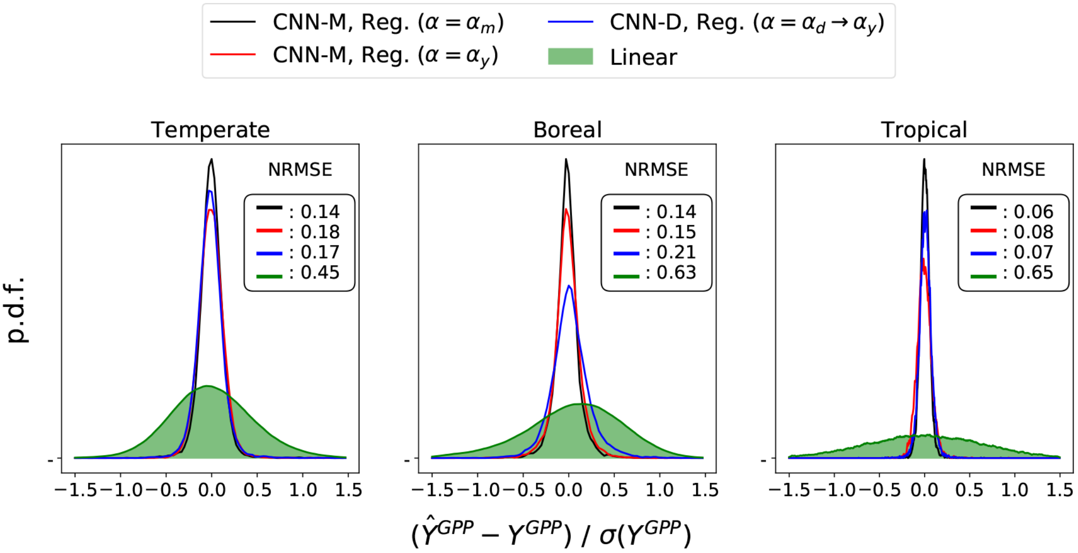

Figure 5. *Distributions of rescaled residuals of yearly gross primary production (GPP) for models in exclusive regression (Reg.) mode, with reported normalized root mean squared error (NRMSE).* $\widehat{Y}^{GPP}$ *indicates the predicted yearly GPP and* Y^{GPP} *its exact value, evaluated for the test set. For CNN-M, when* $\alpha = \alpha_m/\alpha_y$ *the regression task fits monthly/yearly GPP values. For CNN-D,* α *changes during training from* α_d *to* α_y*, fitting daily/yearly GPP values at the beginning/end of training.*

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

tropical site to a large performance improvement of the linear model, showing that the effect of water limitation during the dry season is the main source of nonlinearity for this site.

First, we discuss results from *CNN-M* in regression mode and consider here two limiting values of the meta-parameter α in Equation (1). When $\alpha = 0$, *CNN-M* is trained to minimize only the error of the predicted GPP at the monthly scale. Instead, the model with $\alpha = 1$ minimizes the prediction error directly at the yearly scale, effectively ignoring the monthly GPP values. To use more intuitive notation for *CNN-M*, from now on we will therefore use the notation α_m and α_y referring to $\alpha = 0$ and $\alpha = 1$, respectively. Despite the different ways of leveraging GPP values, the two setups returned comparable distributions of residuals and NRMSEs (Figure 5, black and red lines). The model trained to optimize monthly predictions provides distributions slightly more peaked at zero and lower NRMSEs, but the difference in performance between the two *CNN-M* setups is small. Intermediate values of α have also been explored but did not lead to improved performance (not showed). Regarding the training procedure, the *CNN-M* model fitting intermediate monthly predictions did not require L2 regularization to avoid overfitting, in contrast to the model directly predicting yearly values. Last, in the Supplementary Figures S5–S8, we plot the intermediate monthly predictions given by the former setup, showing a very good correlation between predicted and exact values for all sites.

The distribution of yearly residuals obtained from *CNN-D* is also plotted in Figure 5 (blue lines). In this case, the meta-parameter α balances daily and yearly loss. We found it beneficial to change it along the training. First, the model is trained with $\alpha = 0$, that is with all the loss at the daily scale values. After the model converges, α is changed gradually from zero to one, with the final model having all the loss at the yearly scale. Accordingly, for *CNN-D*, we use the notation α_d and α_y to indicate $\alpha = 0$ and $\alpha = 1$, respectively. The resulting distribution of yearly GPP residuals outperforms the baseline and yearly NRMSE values are slightly larger but comparable to the ones obtained from *CNN-M* models. In the Supplementary Information, we also report the distribution of residuals before switching α from 0 to 1 (Figure S12).

A better understanding of *CNN-D* predictions can be obtained looking at the daily GPP values predicted by the intermediate output layer of *CNN-D*, just after the first training phase with $\alpha = 0$, which is showcased in Figure 6. Note that the mean value of GPP at each day has been subtracted from the time series, that is daily GPP anomalies are reported. While a high day-to-day variability in the GPP time series is observed, one can appreciate that it is followed very closely by the values predicted by the *CNN-D* model. This is also indicated by the values of NRMSEs, which are comparable with the ones obtained from the prediction at the yearly scale.

4.2. Classification

We first present performance of *CNN-M* used in classification mode. The PR curves obtained from the output classification neuron are reported in Figure 7a (green lines). Models in classification mode use only the loss contribution reported in Equation (4) and therefore rely solely on the binary labels and ignore all scalar GPP values. These models show a limited performance with the worst AUC scores, as is particularly evident for the tropical site. We note that L2 regularization was found necessary to train the *CNN-M* architecture in exclusive classification mode, which otherwise would result in strong overfitting and poor performance on the test set.

PR curves substantially improve for all sites when switching to multitasking mode (Figure 7a, red and black lines). This shows the importance of leveraging the additional information present in the GPP time series in order to obtain better models. In the multitasking mode, best results are obtained when the parameter $\alpha = \alpha_m$ is used for the regression loss, that is, when the auxiliary regression task performs optimization of the intermediate monthly GPP predictions. Nevertheless, the difference in AUC scores between models with $\alpha = \alpha_m$ and $\alpha = \alpha_y$ is only marginal, similar to what we found for models in regression mode. For the multitasking mode, regularization was needed only if the auxiliary regression

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

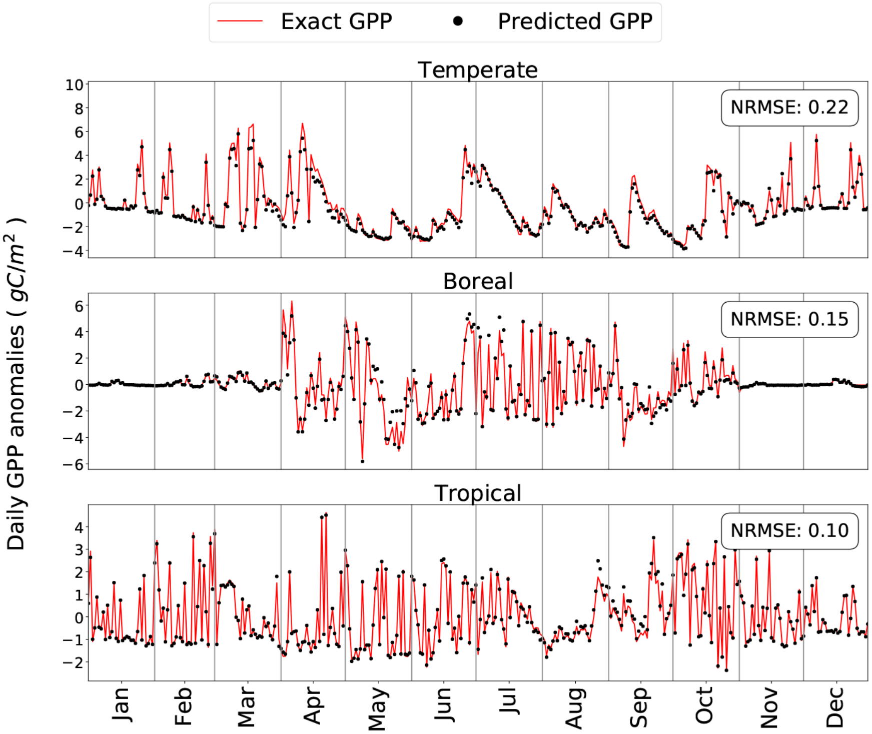

Figure 6. *Daily preditions of GPP from CNN-D (black points), superposed with the exact behavior (red lines), for one random year selected from the test set.*

task consisted in optimizing the predicted yearly GPP value. Instead, when optimizing monthly values, no regularization was needed.

Last, we compare models in classification/multitasking mode against models in regression mode. As discussed in Section 3.2.4, by fixing a threshold on the predicted yearly GPP, it is possible to build PR curves for models in regression mode, even if they do not have in their architecture an explicit classification output neuron, and perform a comparison using the same metrics. PR curves obtained in this way are reported in Figure 7b. Surprisingly, PR curves for models in regression mode improve their AUC scores with respect to classification/multitasking modes, even if they tend to become slightly less smooth. For a given site, however, the difference in classification performance across all models in regression mode can be considered small. Slightly larger AUC scores are obtained with *CNN-M* fitting intermediate monthly values (Figure 7b, black lines). This result is compatible with the previous one obtained from residual analysis (Figure 5). The classifier built on top of *CNN-D* yearly predictions (Figure 7b, blue lines) shows also PR curves with a larger AUC scores than the ones from *CNN-M* in multitasking mode. Performance across sites is consistently best for the tropical site and worse for the boreal site.

In order to better understand why the models trained in regression mode achieve higher AUC scores than the models trained in multitasking mode, we plot the predicted versus exact yearly GPP from a *CNN-M* model trained in multitasking mode (Figure 8). Red points indicate missclassified years according to prediction probability threshold of 0.5. These plots show that the classification neuron is performing predictions coherent with the regression output and is not adding further information. In conclusion, even considering classification metrics related to the distinction of extreme/nonextreme years, the regression approach, focusing its loss function on the prediction of the whole GPP distribution, is superior to the multitasking one.

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

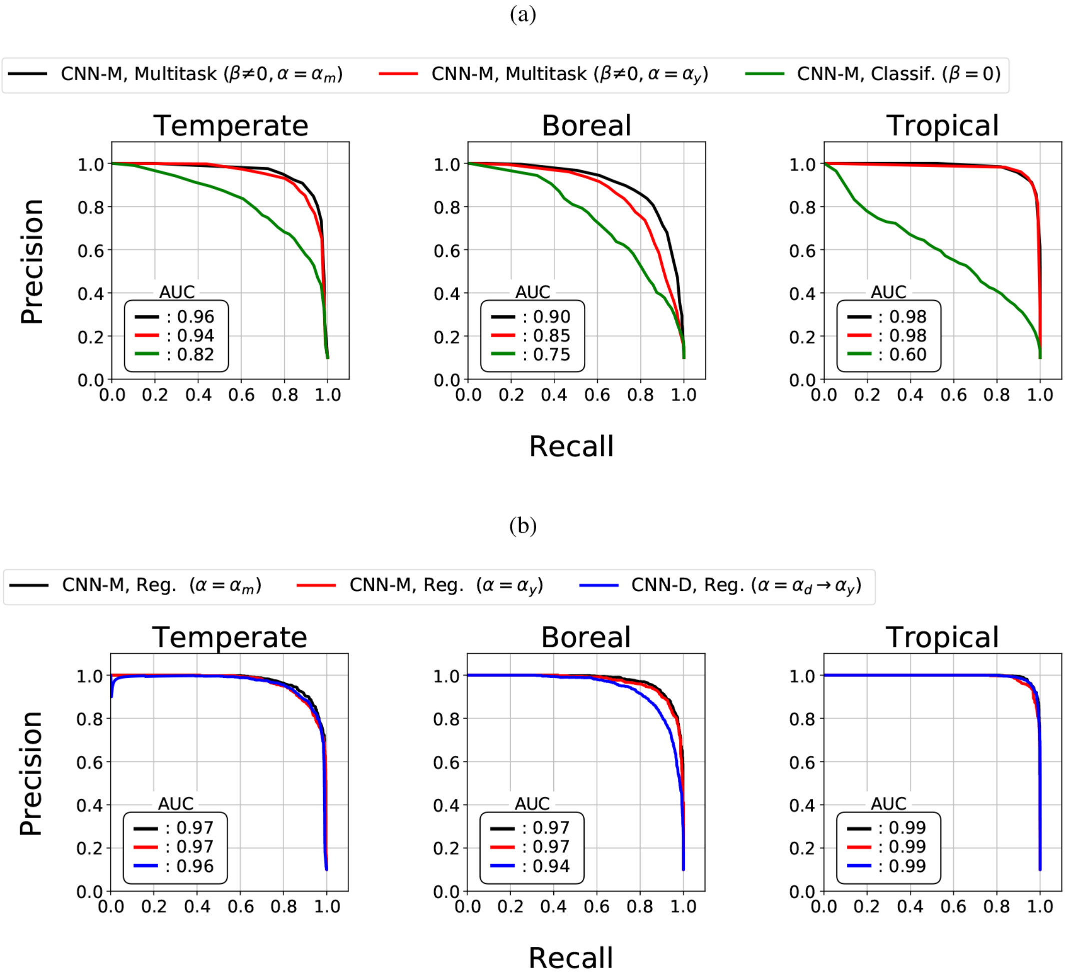

Figure 7. *Precision-recall (PR) curves for different architectures. Corresponding area under curve (AUC) scores are reported on the legends. See caption of Figure 5 and Section 3.2.3 for definition of the different αs. (a) Models based on CNN-M in classification or multitasking mode. Classification is equivalent to multitasking without regression task ($\beta = 0$, see Equation (5)). For these settings the classification output neuron can be directly used to evaluate the PR curves. (b) Models in exclusive regression (Reg.) mode. PR curves are obtained varying a cutoff on the predicted yearly GPP. With this procedure, the PR curves provide a common metric to compare all models.*

5. Discussion

In this work, we demonstrate that one can predict whether a year has low GPP or not from daily meteorological input using ANNs. The different behavior of the trained models in the three sites shows the difficulty of obtaining generalizable meteorological features leading to extremely low values of yearly carbon uptake. According to the type of vegetation growing in each site and the meteorological background conditions, the mapping from daily precipitation, temperature, and radiation to annual gross primary production changes. Our interpretation is that when climate is more variable (e.g., in higher latitudes), changes between different bioclimatic conditions happen more frequently and the the sources of nonlinearities become more abundant and dependent on longer periods. The three sites can be considered representative of the respective climate, that is, tropical, temperate, and boreal. We resorted

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

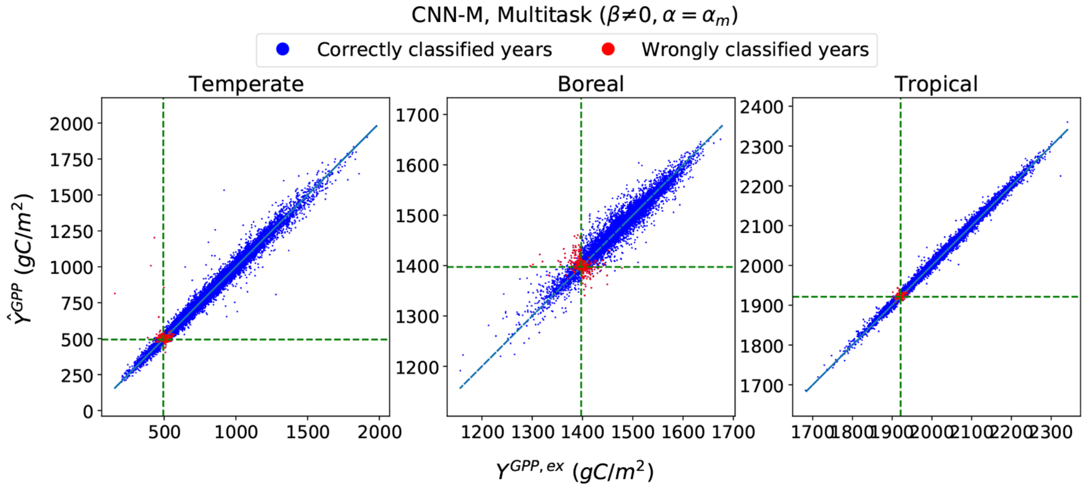

Figure 8. *For the indicated model in multitasking mode, we plot predicted versus exact yearly GPP on the test set. Colors identify correct (blue) and missclassified (red) examples when a fixed threshold of* 0.5 *on the predicted probability is applied. Dashed lines indicate the quantile used to define extreme and normal years.*

to simulated data, which are therefore an idealization of real climatic conditions, but we expect our conclusions to hold qualitatively also in the real world.

The regression results show that all neural network architectures considered here, which leverage the information present in the GPP at different time scales, significantly outperform the yearly predictions of the linear model. *CNN-D* can achieve very good daily GPP predictions as well (Figure 6). Nevertheless, we have shown that yearly residuals obtained from *CNN-D* do not improve over the ones provided by *CNN-M* (Figure 5). Such a result may sound surprising because *CNN-D* exploits 365 GPP values per year to fix parameters during training, whereas *CNN-M* uses 12 or 1 values depending on a monthly ($\alpha = \alpha_m$) or yearly ($\alpha = \alpha_y$) training target, respectively. We hypothesize that near perfect prediction of daily GPP would be required to tackle the cumulation of errors at the annual scale. We further speculate that learning GPP at longer time scales forces the model to neglect potentially irrelevant daily variations and focus on events that are important for the impact of interest. Although we cannot exclude the existence of learning schemes able to better leverage daily GPP values for annual predictions, our results show that this is not a trivial task.

On the other extreme, predicting yearly GPP directly ignoring all intermediate time scales (i.e., the setup provided by *CNN-M* with $\alpha = \alpha_y$) leads already to good residuals, with only one cumulative value of GPP per year used by the network during training. The improvement of NRMSE values (and of residuals distributions) using monthly values (*CNN-M* with $\alpha = \alpha_m$) was non-negligible, but rather small. Nevertheless, only in the latter case *CNN-M* did not require L2 regularization. Therefore, we speculate that the effect of employing the information about the GPP values at a monthly scale has also a regularizing effect on the model, thus enabling the learning of more robust and relevant features.

Finally, in pure classification mode, that is neglecting the GPP distribution and classifying years directly into “extreme” or “normal,” leads to ANN models with the poorest performance, among the ones considered in this work. The amount of data points needed to train a decent model with such a limited information provided to the network depends on the complexity of the dataset as well. In our case (see Figure 7a), we find that models in exclusive classification mode achieve poor AUC score in the tropical site, and intermediate in the temperate and boreal ones. This limited performance can be improved by switching to a multitasking framework, when the model is trained to classify years and to predict values of yearly GPP at the same time. Nevertheless, even these hybrid models achieve lower AUC scores than the

https://doi.org/10.1017/eds.2022.1 Published online by Cambridge University Press

ones in exclusive regression mode, predicting only scalar GPP values (compare Figure 7a and Figure 7b). This finding can be interpreted in the following way. In our framework, the classification task can be considered a subtask of the regression task, because the classification into extreme or normal years is derived from the exact value of the yearly GPP, comparing with the percentile under consideration. Therefore, the multitasking approach can be considered equivalent to a regression approach which focuses its attention on the prediction of GPP in extreme years. It seems the classification tasks finds a "shortcut" that does not generalize well. If the multitasking approach led to AUC scores higher than models based solely on regression, this would have been a sign that the meteorological patterns leading to extremely low GPP values are qualitatively different than the ones driving GPP in normal years. Instead, models in regression mode are able to use the patterns learnt to fit the whole GPP distribution in order to improve classification metrics based solely on the distinction between extreme/nonextreme years. From our results, we speculate therefore that in these datasets similar weather patterns drive GPP variations across all years, either extreme or nonextreme. This hypothesis will be investigated in future works.

In this study, we have used simulated 100,000 years of data, a sample size that is not realistic in observational datasets. Our main goal was to demonstrate the feasibility of the presented approach. Moreover, using novel ideas from the rapidly developing field of explainable machine learning (Samek et al., 2019), we aim to build on these results to generate low-dimensional mappings of weather features to "bad GPP" years, which would allow for a comparison of mappings between well-calibrated process models with real-world observations. Furthermore, very high resolution observations, for instance from satellites (Drusch et al., 2012), would potentially allow a space-for-time substitution to increase the sample size.

6. Conclusions

In this work, we analyzed the performance of several CNN architectures to predict years with extremely low annual GPP from daily meteorological time series. Overall, the employed CNN architectures achieved very good performance. Nonetheless, our results show that the mapping between meteorological data and GPP varies across climatic conditions. In tropical climates, where the source of nonlinearity is due almost exclusively to the presence of a dry season, CNNs show the best performance. We showed that the prediction of years with extremely low carbon uptake using models in exclusive regression mode outperforms models in classification or in mixed (multitasking) mode.

The performance of regression models is relatively stable with respect to the time scale used for intermediate GPP predictions and therefore on the amount of information provided to the network during training. Although daily data were available, the best approach turned out to be the one with an intermediate prediction at the monthly scale, before summing up values to compute yearly GPP. Although we showed that it is possible to obtain very high performance for daily predictions, small errors at the daily scale accumulate to comparably large errors when summing up daily values to obtain yearly GPP. The final error on predicted annual GPP, as well as classification performance into extreme of normal years, was then slightly worse with respect to setups predicting GPP values directly at longer time scales.

Despite the good quality of all the trained models, it remains an open question to understand what are the actual weather features that the CNNs are learning. These learned features are hidden in the trained parameter values. In future work, interpretable machine learning techniques will be used to find out what are the typical meteorological patterns associated with years of extremely low GPP for different climate zones and vegetation types. The acquired understanding of model behavior in this study sets a solid foundation to fully accomplish this task.

Author Contributions. J.Z. designed the project with input from M.V. and S.H. N.P. produced the daily meteorological data. S.L. ran the vegetation model. A.M. and M.V. analyzed the simulations and developed the ANN models. All authors contributed to discussion of the results and writing of the manuscript.

Competing Interests. The authors declare that they have no conflict of interest.

Data Availability Statement. The data that support the findings of this study are openly available in Zenodo at 10.5281/zenodo.5555266.

Funding Statement. This work was funded by the Swiss National Science Foundation (grants no. 189908, 172476 and 179876) and the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX; grant agreement no. VH-NG-1537).

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/eds.2022.1.

References

Bastos A, Fu Z, Ciais P, Friedlingstein P, Sitch S, Pongratz J, Weber U, Reichstein M, Anthoni P, Arneth A, Haverd V, Jain A, Joetzjer E, Knauer J, Lienert S, Loughran T, McGuire PC, Obermeier W, Padrón RS, Shi H, Tian H, Viovy N and Zaehle S (2020) Impacts of extreme summers on European ecosystems: a comparative analysis of 2003, 2010 and 2018. *Philosophical Transactions of the Royal Society B: Biological Sciences 375*(1810), 20190507.

Ben-Ari T, Boé J, Ciais P, Lecerf R, Van Der Velde M and Makowski D (2018) Causes and implications of the unforeseen 2016 extreme yield loss in the breadbasket of France. *Nature Communications 9*(1), 1627.

Drusch M, Del Bello U, Carlier S, Colin O, Fernandez V, Gascon F, Hoersch B, Isola C, Laberinti P, Martimort P, Meygret A, Spoto F, Sy O, Marchese F, and Bargellini P (2012) Sentinel-2: ESA's optical high-resolution mission for GMES operational services. *Remote Sensing of Environment 120*, 25–36

Fatichi S, Ivanov VY and Caporali E (2011) Simulation of future climate scenarios with a weather generator. *Advances in Water Resources 34*(4), 448–467.

Fatichi S, Ivanov VY, Paschalis A, Peleg N, Molnar P, Rimkus S, Kim J, Burlando P and Caporali E (2016) Uncertainty partition challenges the predictability of vital details of climate change. *Earth's Future 4*(5), 240–251.

Flach M, Sippel S, Gans F, Bastos A, Brenning A, Reichstein M and Mahecha MD (2018) Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western Russian heatwave. *Biogeosciences 15*(20), 6067–6085.

Frank D, Reichstein M, Bahn M, Thonicke K, Frank D, Mahecha MD, Smith P, van der Velde M, Vicca S, Babst F, Beer C, Buchmann N, Canadell JG, Ciais P, Cramer W, Ibrom A, Miglietta F, Poulter B, Rammig A, Seneviratne SI, Walz A, Wattenbach M, Zavala MA and Zscheischler J (2015) Effects of climate extremes on the terrestrial carbon cycle: concepts, processes and potential future impacts. *Global Change Biology 21*(8), 2861–2880.

Frieler K, Lange S, Piontek F, Reyer CPO, Schewe J, Warszawski L, Zhao F, Chini L, Denvil S, Emanuel K, Geiger T, Halladay K, Hurtt G, Mengel M, Murakami D, Ostberg S, Popp A, Riva R, Stevanovic M, Suzuki T, Volkholz J, Burke E, Ciais P, Ebi K, Eddy TD, Elliott J, Galbraith E, Gosling SN, Hattermann F, Hickler T, Hinkel J, Hof C, Huber V, Jägermeyr J, Krysanova V, Marcé R, Müller Schmied H, Mouratiadou I, Pierson D, Tittensor DP, Vautard R, van Vliet M, Biber MF, Betts RA, Bodirsky BL, Deryng D, Frolking S, Jones CD, Lotze HK, Lotze-Campen H, Sahajpal R, Thonicke K, Tian H and Yamagata Y (2017) Assessing the impacts of 1.5°C global warming—simulation protocol of the inter-sectoral impact model intercomparison project (ISIMIP2b). *Geoscientific Model Development 10*(12), 4321–4345.

Glorot X, Bordes A and Bengio Y (2011) Deep sparse rectifier neural networks. In Gordon G, Dunson D and Dudík M (eds), *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics, Vol. 15 of Proceedings of Machine Learning Research*. FL, USA: PMLR, pp. 315–323.

He K, Zhang X, Ren S and Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile: IEEE, pp. 1026–1034.

Hersbach H, Bell B, Berrisford P, Hirahara S, Horányi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, De Chiara G, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan RJ, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, de Rosnay P, Rozum I, Vamborg F, Villaume S and Thépaut J-N (2020) The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049.

Hochreiter S and Schmidhuber J (1997) Long short-term memory. *Neural Computation 9*(8), 1735–1780.

Krizhevsky A, Sutskever I and Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In Pereira F, Burges CJC, Bottou L and Weinberger KQ (eds), *Advances in Neural Information Processing Systems 25*. Curran Associates, Inc, pp. 1097–1105.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature 521*(7553), 436–444.

Lenggenhager S, Croci-Maspoli M, Brönnimann S and Martius O (2019) On the dynamical coupling between atmospheric blocks and heavy precipitation events: a discussion of the southern alpine flood in October 2000. *Quarterly Journal of the Royal Meteorological Society 145*(719), 530–545.

Leonard M, Westra S, Phatak A, Lambert M, van den Hurk B, Mcinnes K, Risbey J, Schuster S, Jakob D and Stafford-Smith M (2014) A compound event framework for understanding extreme impacts. *Wiley Interdisciplinary Reviews: Climate Change*, *5*(1), 113–128.

Lienert S and Joos F (2018) A Bayesian ensemble data assimilation to constrain model parameters and land-use carbon emissions. *Biogeosciences 15*(9), 2909–2930.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M and Duchesnay E (2011) Scikit-learn: machine learning in Python. *Journal of Machine Learning Research 12*, 2825–2830.

Peleg N, Molnar P, Burlando P and Fatichi S (2019) Exploring stochastic climate uncertainty in space and time using a gridded hourly weather generator. *Journal of Hydrology 571*, 627–641.

Peleg N, Skinner C, Fatichi S and Molnar P (2020) Temperature effects on the spatial structure of heavy rainfall modify catchment hydro-morphological response. *Earth Surface Dynamics 8*(1), 17–36.

Ronneberger O, Fischer P and Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*. Cham: Springer, pp. 234–241.

Samek W, Montavon G, Vedaldi A, Hansen LK and Müller K-R (2019) *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Vol. 11700. Springer Nature. Cham: Springer.

Sammut C and Webb GI (ed.) (2017) *Encyclopedia of Machine Learning and Data Mining*, 2nd Edn. New York: Springer.

Sippel S, Forkel M, Rammig A, Thonicke K, Flach M, Heimann M, Otto FEL, Reichstein M and Mahecha MD (2017) Contrasting and interacting changes in simulated spring and summer carbon cycle extremes in European ecosystems. *Environmental Research Letters 12*(7), 75006.

Sippel S, Zscheischler J and Reichstein M (2016) Ecosystem impacts of climate extremes crucially depend on the timing. *Proceedings of the National Academy of Sciences 113*(21), 5768–5770.

Sitch S, Smith B, Prentice IC, Arneth A, Bondeau A, Cramer W, Kaplan JO, Levis S, Lucht W, Sykes MT, Thonicke K and Venevsky S (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ dynamic global vegetation model. *Global Change Biology 9*(2), 161–185.

Smith MD (2011) An ecological perspective on extreme climatic events: A synthetic definition and framework to guide future research. *Journal of Ecology 99*(3), 656–663.

Tian H, Yang J, Lu C, Xu R, Canadell JG, Jackson RB, Arneth A, Chang J, Chen G, Ciais P, Gerber S, Ito A, Huang Y, Joos F, Lienert S, Messina P, Olin S, Pan S, Peng C, Saikawa E, Thompson RL, Vuichard N, Winiwarter W, Zaehle S, Zhang B, Zhang K and Zhu Q (2018) The global N2O model intercomparison project. *Bulletin of the American Meteorological Society 99* (6), 1231–1251.

Van der Wiel K, Selten FM, Bintanja R, Blackport R and Screen JA (2020) Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions. *Environmental Research Letters 15*(3), 034050.

Vogel J, Rivoire P, Deidda C, Rahimi L, Sauter CA, Tschumi E, van der Wiel K, Zhang T and Zscheischler J (2021) Identifying meteorological drivers of extreme impacts: an application to simulated crop yields. *Earth System Dynamics 12*(1), 151–172.

Wolf S, Keenan TF, Fisher JB, Baldocchi DD, Desai AR, Richardson AD, Scott RL, Law BE, Litvak ME, Brunsell NA, Peters W and van der Laan-Luijkx IT (2016) Warm spring reduced carbon cycle impact of the 2012 US summer drought. *Proceedings of the National Academy of Sciences 113*(21), 5880–5885.

Zischg AP, Felder G, Weingartner R, Quinn N, Coxon G, Neal J, Freer J and Bates P (2018) Effects of variability in probable maximum precipitation patterns on flood losses. *Hydrology and Earth System Sciences 22*(5), 2759–2773.

Zscheischler J, Fatichi S, Wolf S, Blanken PD, Bohrer G, Clark K, Desai AR, Hollinger D, Keenan T, Novick KA and Seneviratne SI (2016) Short-term favorable weather conditions are an important control of interannual variability in carbon and water fluxes. *Journal of Geophysical Research: Biogeosciences 121*(8), 2186–2198.

Zscheischler J, Martius O, Westra S, Bevacqua ERC, Horton RM, van den Hurk B, AghaKouchak A, Jézéquel A, Mahecha MD, Maraun D, Ramos AM, Ridder N, Thiery W and Vignotto E (2020) A typology of compound weather and climate events. *Nature Reviews Earth and Environment 1*, 333–347.

Zscheischler J, Westra S, van den Hurk B, Seneviratne SI, Ward PJ, Pitman A, AghaKouchak A, Bresch DN, Leonard M, Wahl T and Zhang X (2018) Future climate risk from compound events. *Nature Climate Change 8*, 469–477.

Cite this article: Marcolongo A. Vladymyrov M. Lienert S. Peleg N. Haug S. and Zscheischler J. (2022). Predicting years with extremely low gross primary production from daily weather data using Convolutional Neural Networks. *Environmental Data Science*, 1: e2. doi:10.1017/eds.2022.1