https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

Machine learning and Global Vegetation: Random Forests for Downscaling and Gapfilling

Barry van Jaarsveld[1], Sandra M. Hauswirth[1], and Niko Wanders[1]

[1]Utrecht University, Department of Physical Geography, Princetonlaan 8a, Utrecht, The Netherlands

Correspondence: Barry van Jaarsveld (a.s.vanjaarsveld@uu.nl)

Abstract. Drought is a devastating natural disaster, where water shortage often manifests itself in the health of vegetation. Unfortunately, it is difficult to obtain high-resolution vegetation drought impact, which is spatially and temporally consistent. While remotely sensed products can provide part of this information, they often suffer from data gaps and limitations in spatial or temporal resolutions. A persistent feature among remote sensing products is tradeoffs between spatial resolution and revisiting times, where high temporal resolution is met by coarse spatial resolution and *vice verse*. Machine learning methods have been successfully applied in a wide range of remote sensing and hydrological studies. However, global applications to resolve drought impacts on vegetation dynamics still need to be made available, while there is significant potential for such a product to aid improved drought impact monitoring. To this end, this study predicted global vegetation dynamics based on the Enhanced Vegetation Index (*evi*) and the popular Random Forest algorithm (RF) at 0.1°. We assessed the applicability of RF as a gap filling and downscaling tool to generate spatial and temporal consistent global *evi* estimates. To do this, we trained an RF regressor with 0.1° *evi* data using a host of features indicative of water and energy balances experienced by vegetation and we evaluated the performance of this new product. Next, to test whether the RF is robust in terms of spatial resolution, we downscale global *evi*, the model trained on 0.1° data is used to predict *evi* at 0.01° resolution. The results show that the RF can capture global *evi* dynamics at both the 0.1° (RMSE: 0.02 - 0.4) and at the finer 0.01° (RMSE: 0.04 - 0.6) resolution. Overall errors were higher in the down-scaled 0.01° compared to the 0.1° product. Yet, relative increases remained small, thus demonstrating that RF can be used to create downscaled and temporally consistent *evi* products. Additional error analysis reveals that errors vary spatiotemporally, with underrepresented landcover types and periods of extreme vegetation conditions having the highest errors. Finally, this model is used to produce global spatially continuous *evi* products at both the 0.1° and 0.01° spatial resolution for 2003-2013 at an 8-day frequency.

1 Introduction

Drought is one of the most disruptive natural hazards, causing negative repercussions on the economy, society and the environment, which can affect large areas and populations (Naumann et al., 2014; Vereinte Nationen, 2021). Drought is a complex multivariate phenomenon and its universal identification, and the definition of drought impacts remain a challenge (Blauhut et al., 2016; Vogt et al., 2018; Sutanto et al., 2019). Currently, our understanding of direct drought impacts exceeds those manifested through indirect pathways, and this is primarily due to difficulties in identifying and monitoring indirect drought

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

effects (Vereinte Nationen, 2021). Remotely sensed products that monitor earth system responses during periods of drought are one promising tool that strives towards the alleviation of this issue (AghaKouchak et al., 2015). However, they suffer from tradeoffs between spatial and temporal resolution. The production of high-resolution products will provide a more holistic view of drought responses.

Vegetation is involved in numerous drought-impact pathways (Zhang et al., 2021b). Drought disrupts terrestrial water and carbon cycles, which can reduce the integrity of the ecosystem and associated ecosystem services (Banerjee et al., 2013; Crausbay et al., 2017; Han et al., 2018; Smith et al., 2020). More subtly, vegetation also affects the dynamics of drought propagation; under favourable antecedent conditions, vegetation overshoot may exacerbate and facilitate the onset of rapid and intense droughts (Zhang et al., 2021b). Vegetation is also expected to play a crucial role in shaping drought resistance under future climate change (Vereinte Nationen, 2021). In the absence of such resistance, interventions to alleviate the negative impacts of disrupted ecosystem services can cost up to a billion dollars per drought event (Banerjee et al., 2013; Cammalleri et al., 2020). In addition to disaster mitigation and management, predicting vegetation-drought dynamics is essential for numerous other applications, such as advances in fundamental ecological theory (Murray et al., 2018; Schwalm et al., 2017; Meza et al., 2020; Chen et al., 2021), studies investigating future change in land use, and climate change interventions (Jiang et al., 2017).

It follows that formulating appropriate responses to drought and alleviating the negative effects of ecosystem disruption during these periods requires accurate predictions. Vegetation indices derived from satellite-based remote sensors, such as the Enhanced Vegetation Index (*evi*), have proven to be an indispensable tool for monitoring vegetation at multiple scales, from the fine scale, such as crop patches (Moussa Kourouma et al., 2021; Sharifi, 2021) to the global scale (Huang et al., 2021; Vicente-Serrano et al., 2010). In recent decades, numerous satellite-based vegetation indices have been developed (Li et al., 2021a). However, a persistent feature among these products is trade-offs between spatial resolution and revisiting times, where high temporal resolution is met by coarse spatial resolution and *vice verse*. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS) captures the entire Earth with a high temporal resolution every 1 to 2 days (Zhao and Duan, 2020) with a maximum resolution of 250 m. Landsat and Sentinel-2 data have a higher spatial resolution, 10 and 30 m, but longer revisiting times of approximately 10 and 5 days, respectively (Zhu, 2017; Li et al., 2021a). Revisiting times for Landsat and Sentinel-2 are further prolonged when sensors or retrievals are interrupted by bad weather. Besides temporal frequency, temporal coverage is another important consideration. Coarser scale products are associated with older satellites and have more extended temporal coverage than the newer ones; MODIS products reach as far back as 1999 whereas Sentinel-2 products only go back to 2017. The ideal product for monitoring vegetation dynamics would have global coverage, little to no data gaps, and high spatial and temporal resolution.

Machine learning (ML) methods have been used for downscaling and gap-filling purposes in remote sensing products and can be seen as one tool that may lead to the production of high quality remote sensing products (Zhu et al., 2022; Zeng et al., 2013). Furthermore, ML methods have been successfully applied to a wide range of drought-related (Hauswirth et al., 2021; Shamshirband et al., 2020; Tufaner and Özbeyaz, 2020; Shen et al., 2019; Das et al., 2020; Hauswirth et al., 2022) and remotely sensed vegetation studies (Roy, 2021; Li et al., 2021b; Reichstein et al., 2019). Compared to conventional statistical downscaling techniques, ML is considered the superior alternative; given that no strict statistical assumptions are required,

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

complex and non-linear relationships are well captured and provide high precision (Ebrahimy et al., 2021). One ML algorithm that has been widely applied for downscaling and filling gaps in remote sensing data is the Random Forest Regressor (RF) (Zhang et al., 2021a; Fu et al., 2022; Liu et al., 2020; Wang et al., 2022). Using RF to downscale data involves establishing an RF at a coarse scale and predicting targets at finer resolutions by feeding the algorithm with high-resolution auxiliary data (Liu et al., 2020). Gap-filling can also be achieved by relying on the RF to predict values where data is sparse or missing (Wang et al., 2022). These studies have highlighted that ML methods can accurately predict the dynamics of vegetation (Roy, 2021; Gensheimer et al., 2022). However, studies applying ML methods to global vegetation dynamics concerning drought conditions are less prominent (Li et al., 2021b; Zhang et al., 2021b; Chen et al., 2021).

This study aims to improve the current spatial resolution and missing data limitations of remote sensing-based vegetation health products by linking vegetation health with meteorological and hydrological properties using ML methods. By taking advantage of the wealth of atmospheric, hydrological, and land surface data, we aim to produce a vegetation health product with a high spatial and temporal resolution, provides long records, has global coverage and no data gaps. This was done in three steps; first, assess whether ML is an appropriate tool to predict the condition of vegetation on a global scale and act as a gap-filling tool. Second, to determine whether ML can be used to downscale vegetation conditions and predict values at spatial scales finer than those provided by operational remote sensing products. High degrees of transferability between scales allows for further spatial up- or down-scaling of the RF in future applications while still providing robust results. Lastly, to explore how these products can be applied, we investigated how well the RF predicts the vegetation status in different types of land cover and during drought periods.

2 Materials and Methods

The materials and methods are constructed so that each subsection corresponds to one of the objectives. In the first section, we explain how we tested whether ML is an appropriate tool to predict the status of the vegetation. Then, the procedure was used to determine how well ML can be used to predict the status of vegetation on different scales. Lastly, we detail how the product produced in this study can be used to derive insights into global vegetation dynamics, specifically under drought conditions.

2.1 Predicting globally continuous vegetation status at 0.1° using ML

This section of the material and methods describes the data sources and their relevance to vegetation status; thereafter, the procedure used to fit the RF and make subsequent predictions is detailed. Then, the metrics used to evaluate its accuracy are explained.

2.1.1 Data Sources

The data sources (Table 1 and further information in the following subsections) described below were used to construct a 0.1° resolution dataset to train and test the ML model. The data set spans 10 years, from 2003 to 2013. In the cases where the data

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

were not available at 0.1° resolution, the nearest-neighbour interpolation scheme from `xarray` (Hoyer and Hamman, 2017) was used to match the variables to the same spatial resolution.

Vegetation Index - The reference data used in this study is the *evi* index. *evi* data provide the observational benchmark for the training and validation of the ML-based products created in this study. The *evi* can be used as an indicator of overall vegetation status and health, as it is sensitive to chlorophyll content and correlates with primary production, photosynthesis rates, and vegetation physiognomy (Box et al., 1989). Compared to the more widely used Normalized Difference Vegetation Index, *evi* is considered the superior index, as it is less sensitive to atmospheric conditions and saturation effects in areas of dense vegetation (Gao et al., 2000). These data arise from the Moderate Resolution Imaging Spectroradiometer aboard the Terra and Aqua satellites. Sensors aboard Terra and Aqua are identical, and the 16-day composite images from each sensor are released 8-days apart. In this study, Google Earth Engine's python Application Program Interface (Gorelick et al., 2017) was used to access the terra (MOD13A2.006) and aqua (MYD13A2.006) *evi* data. These two products were combined to produce a quasi-eight-day time series (Didan, 2015, 2021).

Feature Variables - Global vegetation type patterns are largely driven by terrestrial water and energy balances (Hawkins et al., 2003). Similarly, the responses of vegetation to drought are regulated, in part, by water and energy availability (Xu et al., 2010). Accordingly, a suite of data indicative of terrestrial water and energy balances was selected as potential input variables. These variables are introduced below, and Table 1 provides an overview.

Meteorology - Hourly data for total precipitation (*tp*), two-meter temperature (*t2m*), volumetric soil moisture layer 1 (*swvl1*), and soil temperature layer 1 (*stl1*) were retrieved from the hourly ERA5-Land Reanalysis product by the European Centre for Medium-Range Weather Forecasts (Muñoz-Sabater et al., 2021). In addition, potential daily evaporation (*pet*) was acquired from Singer et al. (2021), *pet* is calculated following the Penman-Monteith formulation with ERA5-Land as the input data. All meteorological data were resampled to match the 8-day frequency of the *evi* data. *Tp* was aggregated by taking the cumulative sum of the previous 8 days, whereas the remainder of the variables were averaged over 8 day windows.

Drought Indices - Aside from short-term changes in water availability, it is also key to understand the long-term dynamics to identify drought legacy effects on the current vegetation states (Schwalm et al., 2017). To this end, the Standardised Precipitation Index (*spi*) (McKee et al., 1993) and Standardized Precipitation Evapotranspiration Index (*spei*) (Vicente-Serrano et al., 2010) were used to characterise these legacy effects. The *spi* and *spei* were calculated at the 1, 3, 6, 9, 12 and 24-month aggregation lengths. The different lengths of aggregation are related to types of drought: precipitation, soil moisture, and hydrological droughts. Precipitation and soil moisture droughts mostly correlate short-term deficits in soil water (1-3 months), and are important for vegetation with shallow roots; hydrological drought (6-12 months) can be a good proxy for impacts on shrubs, bushes and trees that have deeper roots and are likely to rely on local groundwater for water (12-24 months). In addition, the inclusion of drought indices allows for the characterisation of past climate memory effects on current vegetation growth (Reichstein et al., 2019; Schwalm et al., 2017) associated with past climatic conditions. The equations and steps for calculating *spi* and *spei* are detailed in Appendix A2.

Landcover Types and Topography - Land cover type is an important predictor of vegetation abundance and health (Meza et al., 2020). Here, the Moderate Resolution Imaging Spectroradiometer Yearly Land cover Types (MCD12Q1.006) were

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

retrieved from the Google Earth Engine's Application Program Interface (Friedl, Mark and Sulla-Menashe, Damien, 2019). In this product, landcover types are classified according to the International Geosphere-Biosphere Programme classification scheme. Barren land, deserts, permanent snow and water bodies were masked in all further analyses. It is important to note that the RF was supplied with the remainder 15 unique landcover types; however, these were collapsed into eight broader classifications for brevity and clarity in the results, discussion and visualisations. To capture the variations in water and energy availability attributable to topographic effects, elevation (*elv*) and height from the nearest drainage basin (*hand*) were accessed from MERIT Hydro, a high-resolution global hydrography map (Yamazaki et al., 2019). Lastly, *slope* and *aspect* was calculated from *elv* using the relevant functions in `xarray-spatial` (Hoyer and Hamman, 2017).

Table 1. Target variable (*evi*) and potential features with accompanying units, spatial resolution (Spat. Res.) and temporal resolution (Temp. Res.)

Name	Units	Spat. Res.	Temp. Res.	Reference
Target Variable				
evi	-	0.01°	8 day	Gao et al. (2000)
Feature Variables				
lc	-		Yearly	Friedl, Mark and Sulla-Menashe, Damien (2019)
elv	m			
hnd	m	0.01°	Static	Yamazaki et al. (2019)
aspect	°			
slope	°			
tp	$mm.day^{-1}$			
t2m	°C			Muñoz-Sabater et al. (2021)
swvl1	-		Hourly	
stl1	°C			
pet	$mm.day^{-1}$	0.1°		Singer et al. (2021)
spi1,spi3,...spi24	-	0.1°	Monthly	this study
spei1,spei3,...spei24	-			

evi - Enhanced Vegetation Index, *lc* - Landcover Types, *elv* - Elevation, *hnd* - Height Above Nearest Drainage, *aspect* - Aspect, *slope* - Slope, *tp* - Total Precipitation, *t2m* - Two Meter Temperature, *swvl1* - Volumetric Soil Water Layer level 1, *stl1* - Soil Temperature layer level 1, *pet* - Potential Evapotranspiration, *spi1,... spi24* - Standardized Precipitation Index (1-month, ... 24-month), *spei1,... spei24* - Standardized Precipitation Evapotranspiration (1-month, ... 24-month). Highlighted rows indiciate that features were dropped from further analysis after conducting feature selection prior to model fitting.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

2.1.2 Random Forest Model

Figure 1. The five sequential steps followed during the RF fitting and evaluation.

While an abundance of ML approaches has been used to predict vegetation status, here the Random Forests Regressor (RF) was selected to link meteorology, land cover, topography, and drought inputs to vegetation health. RF is an ensemble method that fits many decision trees on different subsets of data. RF is advantageous given its relatively straightforwarded implementation, ability to incorporate categorical features, ability to easily identify causal links and limited risk of overfitting. The general pipeline used throughout consisted of five sequential steps (Fig. 1). Here, the RF was implemented in Python 3.9 (Rossum and Drake, 2010) under the `scikit_learn` framework (Pedregosa et al., 2011).

Feature Selection - In an attempt to include only relevant data in the ML model, the list of potential variables described in Section 2.1.1 and Table 1 was evaluated for their ability to provide meaningful information during model fitting. A pairwise Spearman rank correlation was calculated between all features to ensure that input data correlated with *evi*. Those variables that exhibited strong correlations were retained in further analysis, whereas variables that experienced weak correlations were excluded. *Aspect* did not exhibit strong correlations with *evi* (Fig. A1). Similarly, *spi* (at all aggregation times) did not correlate strongly with *evi*. In addition, *spi* and were closely correlated with *spei*, *spi* was excluded in favour of *spei* (Fig. A1). *spi* and *aspect* were excluded from further analyses; features that were excluded are highlighted in Table 1.

Pre-processing - Given that the RF algorithm accepts 2-dimensional numeric arrays as input, the 3-dimensional data was processed so that each unique latitude and longitude was associated with a time series of each variable. The single categorical feature (*lc*) was converted to binary numeric. Each unique landcover type is assigned to a new feature, with 1 indicating presence and 0 indicating absence.

Split Strategy and Hyper-parameter optimisation - `HalvingRandomSearchCV` with a 3-fold cross-validation approach was applied to refine the number of estimators and maximum depth. This hyper-parameter optimisation provides the optimal configuration for the RF so that the critical vegetation dynamics are captured while simultaneously reducing the RF complexity and preventing over-fitting. The hyper-parameter optimisation focused on two parameter settings, namely, `Maximum_depth` and the `number_of_estimators`; the search space was 1-40 and 1-20, respectively. Increasing the `Maximum_depth` and `number_of_estimators` past 12 and 15, respectively, yielded only marginal increases in test

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

scores (Fig. 2a). Given that only the risk of overfitting increases with increasing `Maximum_depth` and `number_of_estimators` and only marginal increases in test scores are observed past these points, 12 and 15 were identified as the optimal settings.

After determining optimal parameter settings, the data were split into training and validation sets. However, three-dimensional data could conceivably be split along the temporal dimension where the model is trained on all locations with only a subset of the temporal availability (i.e., temporal splitting), or the data can be split according to location where only a subset of the grid pixels are selected for training but over the entire available period (i.e., spatial splitting). Given that previous research has highlighted that RF performance is sensitive to spatial vs temporal splitting, this is especially true for extreme events such as droughts (Hauswirth et al., 2021). We conducted a cursory analysis to determine whether a temporal or spatial splitting approach better balances trade-offs between computational complexity and learning rates. Learning curves for cursory RF models using each splitting approach were quantified and compared. Each model was supplied with increasing training sizes, and test scores were calculated and plotted to visualise learning curves. This cursory analysis revealed that spatial splitting yields faster learning curves than the temporal splitting approach (Fig. 2b); therefore, spatial splitting was identified as the preferred approach.

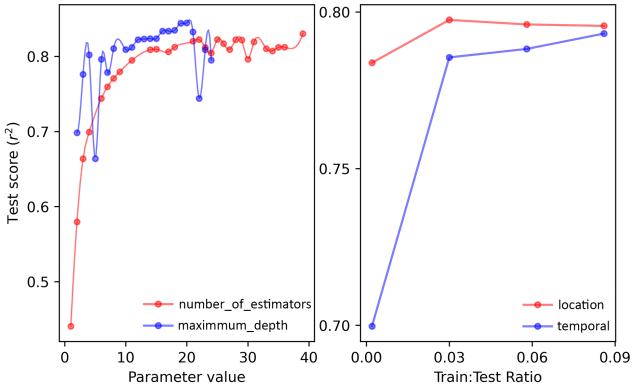


Figure 2. (a) Evolution of RF performance during `HalvingRandomSearchCV` hyper-parameter optimization of: `maximum_depth` (blue) and `number_of_estimators` (red). (b) RF performance following the incremental increase of train set size using a location (red) based split approach compared to a temporal (blue) based split approach.

Train - For the final RF model, a spatial split with a (0.06:0.94) (train: predict) ratio was used to train the final model. A 0.06:0.94 split was chosen, and there was very little increase in performance past training sixes of 6% (Fig. 2b). `Maximum_depth` and `number_of_estimators` were set at 12 and 13, respectively. The parameters that were not subjected to hyper-parameter optimisation were set as follows: the `squared_error` criterion was used to measure the quality of the splits in branches, the maximum `number_of_features` considered in each split was set at `auto`, and the minimum and maximum `samples_per_leaf_nodes` was set at 1 and 2, respectively.

Predict and Evaluate - As a first-pass assessment of overall performance, the model was scored using the validation and the default coefficient of determination (R^2) scorer in the scikit RF implimentation. The model predictions were further evaluated

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

by calculating the root mean squared error (RMSE) and Pearson correlation coefficients. These were calculated independently for each grid cell to provide information on the spatial variation of errors. Last, to gain insight into which features were the most essential for predicting *evi*, global feature importance was calculated using Shapley Additive exPlanations' (SHAP) TreeExplainer (Lundberg et al., 2020).

2.2 Downscaling globally continuous vegetation status using ML

In this section, the focus shifted toward whether RF can be used to downscale global *evi* values, that is, whether a model trained on 0.1° can accurately predict *evi* at a finer 0.01° scale. To this end, a 0.01° data set was compiled. In cases where data were not at the 0.01° resolution, the nearest neighbour interpolation scheme from `xarray` (Hoyer and Hamman, 2017) was used to match the variables to the same spatial resolution. This data set was used as new input data to the already trained RF model to predict *evi* at the 0.01° scale. The evaluation approach for the downscaled values remained much the same, the overall model accuracy was assessed using (R^2) and (RMSE), and Pearson correlation coefficients were calculated for each grid cell.

2.3 Applicability of ML informed vegetation status products during periods of drought

One noticeable shortcoming of the RF is its relatively poor ability to predict extreme values depending on the training selection (Hauswirth et al., 2021). To determine to what degree this may influence the generality of the two products mentioned above, we further investigated the accuracy of the predicted *evi* under low growing conditions by calculating the anomaly correlation index (Eq. 1), where $eviA_{i,j}$ denotes *evi* anomaly for the month j in year i, $\bar{evi}_{,j}$ denotes the average *evi* of month *j* over 2003-2013; σ stands for the standard deviation of *evi* during the period.

$$eviA_{i,j} = \frac{evi_{i,j} - \bar{evi}_j}{\sigma} \tag{1}$$

3 Results

The results here are presented in three parts. First, the results of the model trained on the 0.1° data are presented; here, the focus is retained on the model's performance and ability to predict the status of the vegetation at the spatial resolution it is trained. We also touch on which features are most important in predicting the status of the vegetation. Subsequently, we present the model's performance when used to downscale *evi* and predict 0.01° data. Lastly, we explore how this module can be used to gain insight into global vegetation dynamics by assessing the accuracy of both products under drought conditions.

3.1 Random Forest Regressor at the 0.1° resolution

SHapley Additive exPlanations values provided an understanding of the relative importance of each feature in predicting *evi*. The most important features were those associated with meteorology, landcover type and topography; drought indices proved

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

to be a lower degree of information (Fig. 3). The model was able to reproduce global vegetation patterns by correctly predicting high vegetation density in tropical forests and low vegetation density in arid and urban regions of the world (Fig. 4).

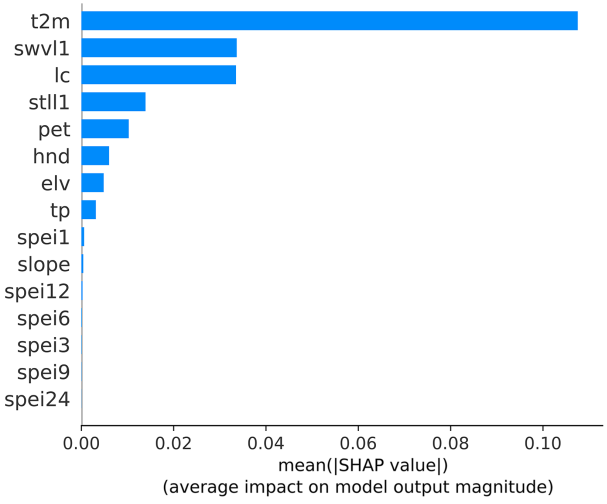


Figure 3. Feature importance for the RF predicting *evi* at 0.1°. The features are ordered by level of importance, with higher mean SHAP values indicating higher importance.

When trained on only 6% of the data, the RF was able to predict global *evi* accurately with a spatial resolution of 0.1° (R^2 = 0.86; Fig. 4, 5, 6 & 7a). Looking more closely at the distribution of errors, less than 1% of grid cells showed negative correlations and more than 80% showed correlations higher than 0.5 (Fig. 7c) and RMSE ranged between 0.02 and 0.4 (mean: 0.05 ± 0.03; Fig. 7d). However, it is important to note that the accuracy was neither spatially nor temporally uniform. Landcover types were an important feature in determining predictive ability. The predictions of *evi* in areas dominated by urban and crop landcover types showed the highest degree of error (Fig. 5a & 6a). On the contrary, the most natural types of land cover, such as forests and grasslands, were the most accurately represented by the model (Fig. 5a & 6a). For all types of land cover, the periods of maximum and minimum *evi* were less accurately predicted than the intermediate periods (Fig. 6a).

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

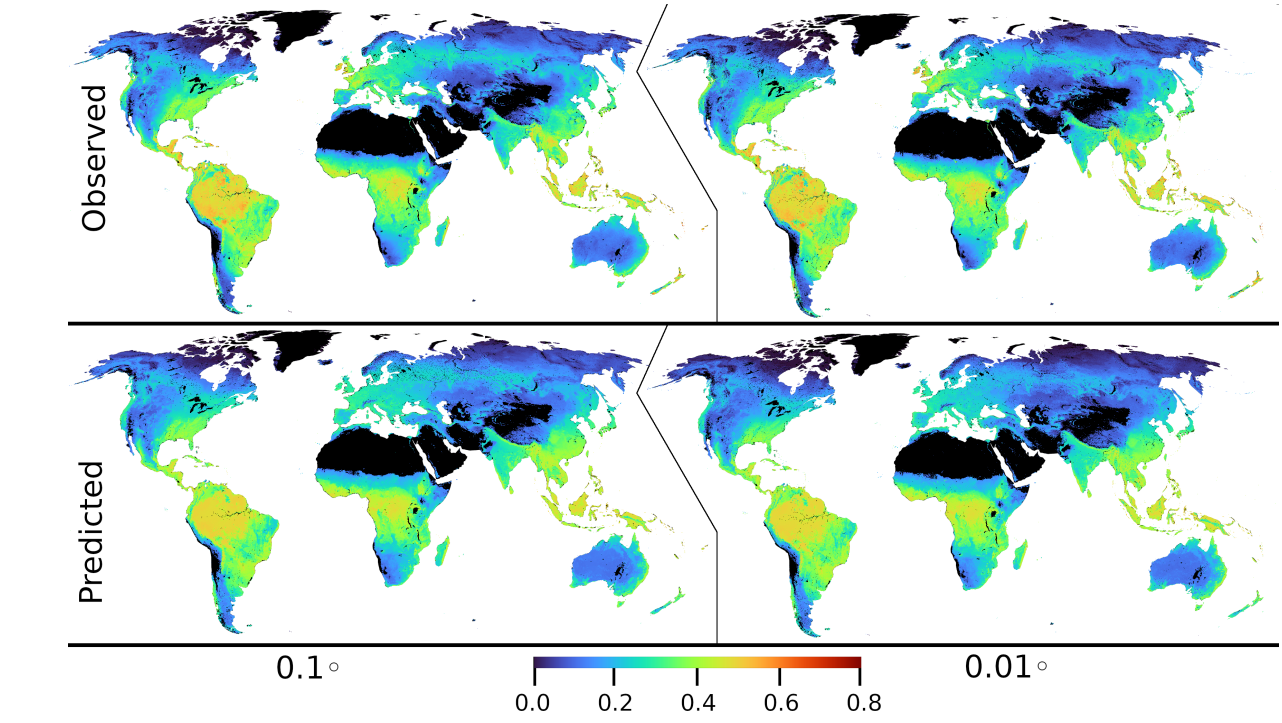


Figure 4. Mean (2003 - 2013) observed (top) and mean predicted *evi* values by the model at the 0.1° (left) and 0.01° (right). Barren land, deserts, permanent snow, and water bodies were masked and represented by black.

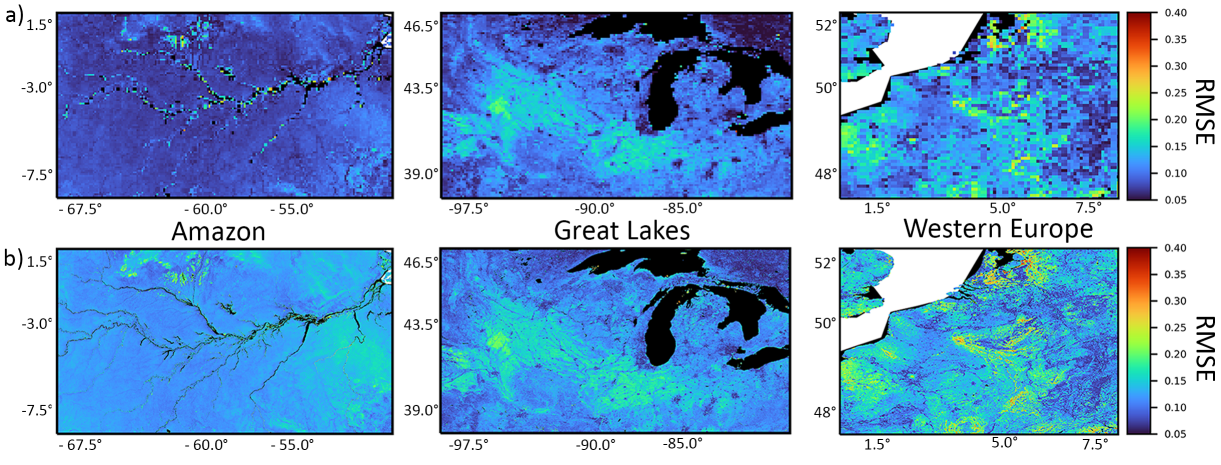
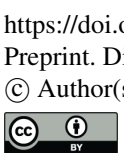


Figure 5. RMSE calculated for the Amazon Basin, Great Lakes and Western Europe for predicted *evi* values by the model at the (a) 0.1° and (b) 0.01°.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023



Hydrology and Earth System Sciences
Discussions

3.2 Random Forest Regressor accuracy at the 0.01° resolution

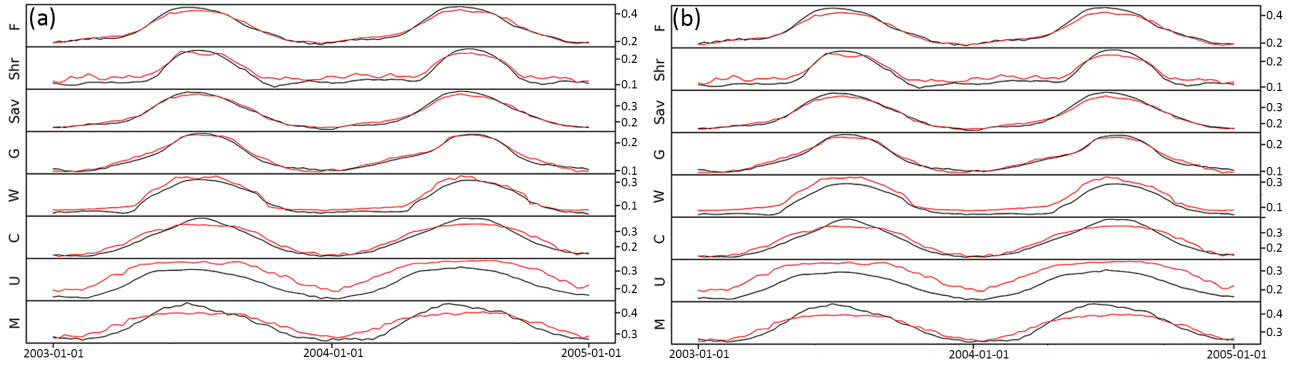


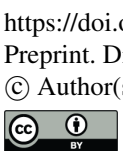
Figure 6. Time series of average predicted (blue) and observed (red) *evi*, per major land cover type at (a) 0.1° and (b) 0.01°. F=Forest, Shr=Shrubland, Sav=Savanna, G=Grassland, W=Wetlands, C=Crops, U=Urban, M=Mixed.

When the model trained with 0.1° data was used to predict *evi* at 0.01° spatial resolution, there was a slight drop in accuracy. The predictive capacity was still good but reduced compared to the 0.1° product, with a median R^2 of 0.75 (Fig. 7b). Again, the RF was able to accurately capture spatial and temporal vegetation dynamics when supplied with 0.01° data (Fig. 4 & 6b). The errors also increased, the proportion of grid cells displaying negative correlations now being 5% (Fig. 7c) compared to less than 1% for the 0.1°product. RMSE ranged between 0.04 and 0.6 (mean: 0.09 $\pm$ 0.07; Fig. 7d), with the majority of the grid cells exhibiting RMSE around 0.05.

3.3 Accuracy under drought conditions

The ACC analysis revealed that the RF was still able to capture *evi* anomalies (Fig. 8), but to a lesser extent compared to overall performance (Fig. 7c). The majority of grid cells showed positive correlations, with less than 10% displaying negative correlations. At least 50% of grid cells exhibited an ACC of 0.25 for 0.1° compared to 45% when *evi* was predicted at 0.01° (Fig. 8). This indicates that for that 90% of the locations, the RF can reproduce anomalies from the average seasonal cycle and thus can be used to identify periods of negative or positive *evi* impacts resulting from droughts or more favourable growing conditions.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023



Hydrology and Earth System Sciences
Discussions

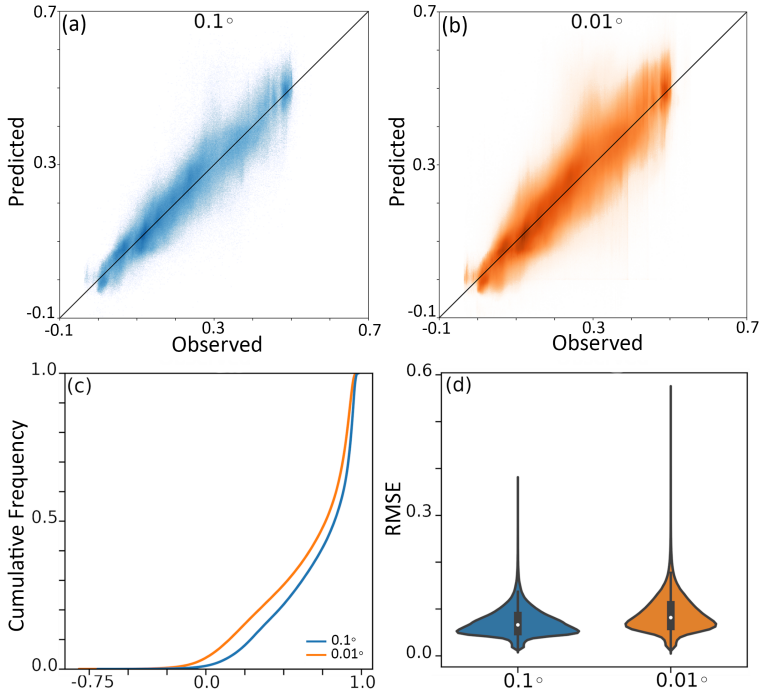


Figure 7. : (a) Scatter plot of observed and predicted evi at 0.1° and (b) 0.01°; Cumulative distribution function for (c) Pearson Correlation Coefficients overall grid pointsat 0.1° (blue) and 0.01° (orange), (d) violin plot of RMSE for all grid points at 0.1° and 0.01°

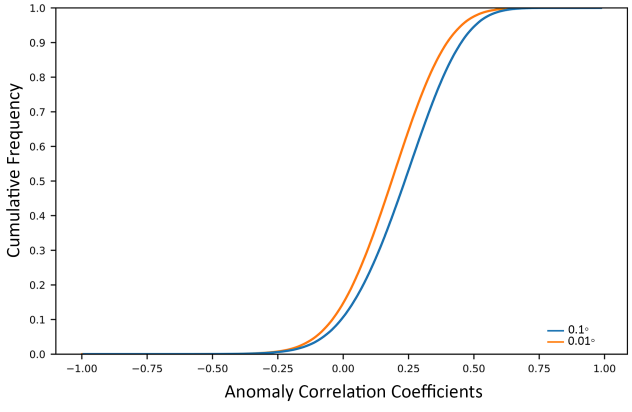


Figure 8. Cumulative distribution curves of anomaly correlation coefficients for *evi* predicted by a RF at 0.1° (blue) and 0.01° (orange).

4 Discussion

240 This study assessed whether the RF algorithm is an appropriate tool for predicting *evi* dynamics at the global scale. RF was evaluated for its ability to be used as a gap-filling and downs-scaling tool. The discussion is outlined as follows, first, the

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

overall performance of the RF is discussed; after that its usefulness as a gap-filling or downscaling tool is critically evaluated. We then highlight important outcomes from applying the model during periods of drought. Lastly, we discuss the importance of landcover types concerning model accuracy.

4.1 Overall performance

The RF successfully predicted *evi* at 0.1° scale from meteorological, topography, and landcover types as input data. Of these data, features related to water and energy balances were most important in predicting *evi*. The RF successfully captured annual vegetation growth cycles and was able to distinguish between the main global biomass with high accuracy. These results add to the current body of research showing that the RF is a powerful technique for predicting temporal and spatial vegetation dynamics from remote sensor data (Roy, 2021; Staben et al., 2018; Wang et al., 2021). Error analysis revealed that prediction accuracy could have been more homogeneous across space and time and varied according to the growing season and land cover type. This behaviour can also be linked to the relative abundance of land cover types, where more dominant land cover types are simulated with higher accuracy.

4.2 RF for Gap-filling

A promising aspect of this study is that the RF can accurately predict *evi* at unseen geographic locations when trained on relatively few data - only 6% in this case. It follows that this approach can be used to for gap-filling purposes and produce high-resolution vegetation indices from other satellite sources or be used in conjunction with satellite products. For example, Landsat and Sentinel-2 data produce high-resolution data vegetation products; however, retrievals are strongly affected by weather conditions, which results in data gaps. In addition, its relatively low orbiting altitude means that the spatial coverage for each pass over is low. The approach outlined in this study could be applied to Landsat and Sentinel-2, to produce continuous vegetation index data sets at the 30-10m spatial resolution. This approach has been previously used to impute missing values for other remote sensing products like land cover types (Holloway-Brown et al., 2021), leaf traits (Moreno-Martínez et al., 2018) and soil moisture (Nguyen et al., 2022) and can now be the extent to vegetation health or *evi*.

4.3 RF as downscaling tool

The RF accurately predicted *evi* at finer spatial scales than was trained, successfully predicting *evi* at a scale of 0.01° using high-resolution auxiliary data. However, it should be noted that this resulted in a reduction in precision compared to the 0.1°product. This is an expected result, given that *evi* at the 0.01° resolution will exhibit greater variances and more extreme values during periods of high and low growth. Scale-dependent drivers of vegetation dynamics may be another phenomenon that contributes to decreased precision when predicting *evi* at the 0.01° using a model trained at a coarser resolution. Meteorology has been shown to be tightly coupled to vegetation at the ecosystem scale but less so at finer scales, where biotic processes, such as competition, herb ivory, disease, and fire, are more important (Franklin et al., 2020). When predicting *evi*, the relative increases in error remained small. This product can be of particular use in cases where the benefits of having high resolution long-term

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

evi products outweigh the limitations associated with error increases. The product presented here can and should be used in further studies investigating global vegetation dynamics.

4.4 RF for predicting drought effects

Compared to the overall performance, the RF was less capable of capturing extreme values of *evi*. Increased error among extreme values is a known limitation of the RF (Hauswirth et al., 2021). During RF training, an evaluation metric, in this case `squared_error`, is used to minimise the error for the model as a whole. In this scenario, optimal fits inevitably result in reduced errors for values close to the mean at the expense of inflated errors for the outliers (Ribeiro and Moniz, 2020). In the current study, this means that *evi* during normal growth periods is prioritised over periods of extremely low or high vegetation growth. The production of ML frameworks that accurately reproduce vegetation responses during extreme periods is an essential consideration for future research directives.

4.5 Importance of Landcover Types

Varying error according to landcover type in the 0.1° and 0.01° is expected for at least two reasons. The first relates to the inherent features of the RF algorithm itself, and the second to the environmental process that affects the dynamics of *evi*. A limitation of the RF algorithm is that when data is imbalanced, underrepresented groups are less well explained by the algorithm. Accordingly, accuracy varied according to a proportional abundance of landcover types (Jung et al., 2020). Dominant landcover types, such as forests and grasslands, displayed the least amount of error; in contrast, minority landcover types regions that have undergone human modification (i.e., urban areas and croplands) were associated with the highest error. Second, the features used in this study may not incorporate processes critical to vegetation status equally among landcover types (Moussa Kourouma et al., 2021). Forests, grasslands, and other natural ecosystems are closely coupled to natural weather processes. However, croplands and urban areas may be less influenced by weather and more influenced by anthropogenic manipulations of water and energy balances (Zhang et al., 2004; Hawkins et al., 2003; Tang et al., 2021). A potential solution to this problem is to rely on Extreme Gradient Boosted Decision Trees, which have been shown to provide more accurate predictions where data are imbalanced (Li et al., 2021b) or include information on human-water management to better represent drought responses (Wanders and Wada, 2015).

Landcover-specific variations in the model's ability to predict vegetation are an important outcome of this study. Apart from the statistical reasons detailed in the previous paragraph as potential mechanisms for this phenomenon, an additional, and most likely compounding explanation is that the data used to predict *evi* may be more relevant for some landcover types and levels of vegetation growth than others. For instance, vegetation status in urban areas and croplands shows weak correlation or high errors (Fig. 5 & 6). The meteorological data used here to predict *evi* may not be the only factor driving the vegetation dynamics in human-modified areas. It is possible that irrigation and water management influence vegetation. Indeed, vegetation in urban areas have been shown to grow more rapidly and have a longer growing season than rural counterparts; this is thought to be driven by higher temperatures, high concentration of airborne phosphorous and other aerosol pollutants (Sicard et al., 2018a, b; Pretzsch et al., 2017). In contrast, natural forests and grasslands show high levels of accuracy and correlations, thus suggesting

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

that the data used here is appropriate for the machine learning models to capture vegetation dynamics. Similarly, poor accuracy in wetlands is not unexpected as wetland vegetation is primarily driven by water quality, salinity, and pH (Grieger et al., 2021). On the contrary, forests and grasslands show high accuracy when using meteorological variables, since these are important drivers of vegetation growth in these areas. Although not directly related to vegetation, Hauswirth et al. (2021) showed that by including water management practices in machine learning models, the predictions of groundwater head and stream flow were more accurately predicted. Therefore, we further iterate the need to provide models with appropriate input data sources.

5 Conclusions

The results from this study reveal that the RF is an appropriate method for predicting *evi* at the global scale, at the 0.1° and downscaled 0.01° resolution. In general, RF was capable of predicting *evi* dynamics with high accuracy; global patterns of vegetation and temporal dynamics were well captured. However, it is essential to note that higher errors were associated with underrepresented landcover types and periods of extreme vegetation growth, such a drought periods. Lower accuracy for underrepresented classes in unbalanced data sets and a hampered ability to predict extreme values is a common criticism of the RF. In accordance with this study, landcover types that account for a smaller fractional cover of the earth's surface, and periods of extreme vegetation growth, were associated with the highest error. Predicting *evi* at a finer resolution resulted in increased errors. This is attributable to higher variances in the 0.01° product compared to 0.1° and it is important to note that the relative increases remained small.

The results here also highlight the use of RF for efficiently and accurately predicting missing data and downscaling, ultimately allowing for the production of spatially continuous *evi* data at very high spatial and temporal resolutions. Toward this end, this study produces spatially continuous *evi* product at 0.1° and 0.01° resolution, which could be used to fill existing gaps in satellite observations or in conjunction with satellite data to have improved monitoring of drought impacts on vegetation.

This study adds to previous research efforts that have successfully applied the RF in predicting vegetation status. Here the RF was used to produce a global spatial and temporally continuous *evi* product at 0.1° and 0.01°, with a median R^2 of 0.86 & 0.75, respectively. The approach outlined in this study could be applied to Landsat and Sentinel-2, to produce continuous vegetation index data sets at the 30-10m spatial resolution. The RF algorithm is a powerful technique for predicting temporal and spatial vegetation dynamics from remote sensor data, as well as those using RF for gap filling purposes on remote sensing products. The novelty of this product, compared to previous studies, is that it has global coverage, high spatial and high temporal resolution.

Author contributions. **BvJ:** Data curation, Formal analysis, Writing – review & editing.**SH:** Data curation, Formal analysis, Writing – review & editing. **NW** Conceptualization, Formal analysis, Writing – review & editing.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

335 *Competing interests.* NW is a member of the editorial board of journal Hydrology and Earth System Sciences. The peer-review process was guided by an independent editor, and the authors have also no other competing interests to declare.

Acknowledgements. Steven M. de Jong is thanked for his valuable input on previous versions of this manuscript. SH acknowledges funding from the Cooperate Innovation Program and the Department of Water, Transport and Environment at the Dutch National Water Authority, Rijkswaterstaat. NW acknowledges funding from NWO 016.Veni.181.049.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Appendix: A1. Feature Selection

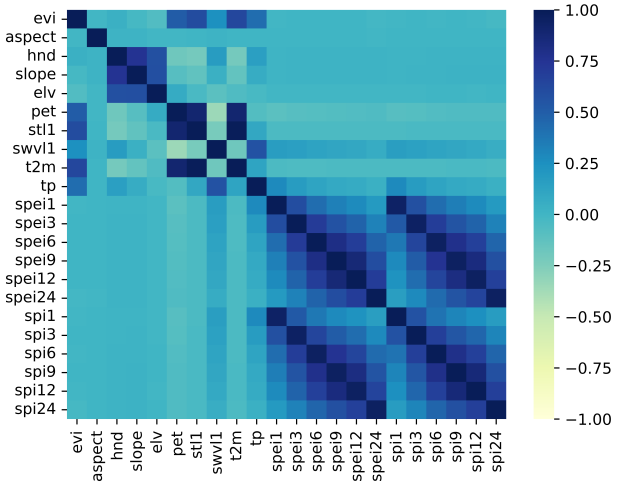


Figure A1. Correlation Matrix of pairwise Spearman rank correlation coefficients between all potential variables

Appendix: A2. Drought Indices Calculations

For the calculation of *spi*:

$$x = \sum_{i}^{m} tp_i \tag{A1}$$

where *i* is the month in question and $m = i - scale$.

For the calculation of *spei*:

$$x = \sum_{i}^{m} D_i \tag{A2}$$

where: $D_i = tp_i - pet_i$ and

$$x_{i,j}^{k} = \begin{cases} \sum_{l=13-k+j}^{12} tp_{i-j,l} + \sum_{l=1}^{j} tp_{i,l}, & \text{if } j < k \\ \sum_{l=j-k+1}^{j} tp_{i,l}, & \text{if } j \geq k \end{cases} \tag{A3}$$

This time series is then fitted to a gamma distribution taken the following steps:

First α and β fitting parameters as calculated as:

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023


Hydrology and Earth System Sciences
Discussions

$$\hat{\alpha} = \frac{1}{4A}(1 + \sqrt{1 + \frac{4A}{3}}) \tag{A4}$$

Where $A = \ln(\bar{x}) - \frac{\sum \ln(x)}{n}$ with n being number of observations.

$$\hat{\beta} = \frac{\bar{x}}{\alpha} \tag{A5}$$

The gamma distribution probability density (Eq. A6) function with respect to *x* and including the calculated estimates for α and β can be inserted to produce an equation for the cumulative probability of a value for (Eq. A7).

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{\frac{x}{\beta}} \tag{A6}$$

where α is the shape parameter and β is the scale parameter and $\Gamma(a) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$

$$G(x) = \frac{1}{\hat{\beta}^{\hat{\alpha}}\Gamma(\hat{\alpha})} \int_0^{x} x^{\hat{\alpha}} e^{\frac{-x}{\hat{\beta}}} dx \tag{A7}$$

then substituting t for $\frac{x}{\hat{\beta}}$ results in the incomplete gamma distribution (Eq. A8)

$$G(x) = \frac{1}{\Gamma(\hat{\alpha})} \int_0^{x} t^{\hat{\alpha}-1} e^{-1} dt \tag{A8}$$

Values of the incomplete gamma function can be computed using Eq. A9

$$H(x) = q + (1 - q)G(x) \tag{A9}$$

Finally, values computed from Eq. A9 are transformed into the standard normal distribution to yield the *spi* and *spei* at the relevant time scales. These calculations were completed using the relevant algorithms in the `climate_indices` python package (Adams, 2021) using *tp*, *pet*, and *t2m* detailed in Section 2.1.2.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

References

Adams, J.: Climate Indices in Python, https://github.com/monocongo/climate_indices, original-date: 2017-06-13T15:21:07Z, 2021.

AghaKouchak, A., Farahmand, A., Melton, F. S., Teixeira, J., Anderson, M. C., Wardlow, B. D., and Hain, C. R.: Remote sensing of drought: Progress, challenges and opportunities: REMOTE SENSING OF DROUGHT, Reviews of Geophysics, 53, 452–480, https://doi.org/10.1002/2014RG000456, 2015.

Banerjee, O., Bark, R., Connor, J., and Crossman, N. D.: An ecosystem services approach to estimating economic losses associated with drought, Ecological Economics, 91, 19–27, https://doi.org/10.1016/j.ecolecon.2013.03.022, 2013.

Blauhut, V., Stahl, K., Stagge, J. H., Tallaksen, L. M., De Stefano, L., and Vogt, J.: Estimating drought risk across Europe from reported drought impacts, drought indices, and vulnerability factors, Hydrology and Earth System Sciences, 20, 2779–2800, https://doi.org/10.5194/hess-20-2779-2016, 2016.

Box, E. O., Holben, B. N., and Kalb, V.: Accuracy of the AVHRR vegetation index as a predictor of biomass, primary productivity and net CO 2 flux, Vegetatio, 80, 71–89, publisher: Springer, 1989.

Cammalleri, C., Naumann, G., Mentaschi, L., Formetta, G., Forzieri, G., Gosling, S., Bisselink, B., De Roo, A., and Feyen, L.: Global warming and drought impacts in the EU: JRC PESETA IV project : Task 7., Publications Office, LU, https://data.europa.eu/doi/10.2760/597045, 2020.

Chen, Z., Liu, H., Xu, C., Wu, X., Liang, B., Cao, J., and Chen, D.: Modeling vegetation greenness and its climate sensitivity with deep-learning technology, Ecology and Evolution, 11, 7335–7345, https://doi.org/10.1002/ece3.7564, 2021.

Crausbay, S. D., Ramirez, A. R., Carter, S. L., Cross, M. S., Hall, K. R., Bathke, D. J., Betancourt, J. L., Colt, S., Cravens, A. E., Dalton, M. S., Dunham, J. B., Hay, L. E., Hayes, M. J., McEvoy, J., McNutt, C. A., Moritz, M. A., Nislow, K. H., Raheem, N., and Sanford, T.: Defining Ecological Drought for the Twenty-First Century, Bulletin of the American Meteorological Society, 98, 2543–2550, https://doi.org/10.1175/BAMS-D-16-0292.1, 2017.

Das, P., Naganna, S. R., Deka, P. C., and Pushparaj, J.: Hybrid wavelet packet machine learning approaches for drought modeling, Environmental Earth Sciences, 79, 221, https://doi.org/10.1007/s12665-020-08971-y, 2020.

Didan, K.: MOD13A2 MODIS/Terra Vegetation Indices 16-Day L3 Global 1km SIN Grid V006 [Data set]. NASA EOSDIS LP DAAC, 2015.

Didan, K.: MODIS/Aqua Vegetation Indices Monthly L3 Global 0.05 Deg CMG V061,[data set], NASA EOSDIS Land Processes DAAC, 2021.

Ebrahimy, H., Aghighi, H., Azadbakht, M., Amani, M., Mahdavi, S., and Matkan, A. A.: Downscaling MODIS Land Surface Temperature Product Using an Adaptive Random Forest Regression Method and Google Earth Engine for a 19-Years Spatiotemporal Trend Analysis Over Iran, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 2103–2112, https://doi.org/10.1109/JSTARS.2021.3051422, 2021.

Franklin, O., Harrison, S. P., Dewar, R., Farrior, C. E., Brännström, , Dieckmann, U., Pietsch, S., Falster, D., Cramer, W., Loreau, M., Wang, H., Mäkelä, A., Rebel, K. T., Meron, E., Schymanski, S. J., Rovenskaya, E., Stocker, B. D., Zaehle, S., Manzoni, S., van Oijen, M., Wright, I. J., Ciais, P., van Bodegom, P. M., Peñuelas, J., Hofhansl, F., Terrer, C., Soudzilovskaia, N. A., Midgley, G., and Prentice, I. C.: Organizing principles for vegetation dynamics, Nature Plants, 6, 444–453, https://doi.org/10.1038/s41477-020-0655-x, 2020.

Friedl, Mark and Sulla-Menashe, Damien: MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006, https://doi.org/10.5067/MODIS/MCD12Q1.006, type: dataset, 2019.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Fu, R., Chen, R., Wang, C., Chen, X., Gu, H., Wang, C., Xu, B., Liu, G., and Yin, G.: Generating High-Resolution and Long-Term SPEI Dataset over Southwest China through Downscaling EEAD Product by Machine Learning, Remote Sensing, 14, 1662, https://doi.org/10.3390/rs14071662, 2022.

Gao, X., Huete, A. R., Ni, W., and Miura, T.: Optical–Biophysical Relationships of Vegetation Spectra without Background Contamination, Remote Sensing of Environment, 74, 609–620, https://doi.org/10.1016/S0034-4257(00)00150-4, 2000.

Gensheimer, J., Turner, A. J., Köhler, P., Frankenberg, C., and Chen, J.: A convolutional neural network for spatial downscaling of satellite-based solar-induced chlorophyll fluorescence (SIFnet), Biogeosciences, 19, 1777–1793, https://doi.org/10.5194/bg-19-1777-2022, 2022.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., and Moore, R.: Google Earth Engine: Planetary-scale geospatial analysis for everyone, Remote Sensing of Environment, 202, 18–27, https://doi.org/https://doi.org/10.1016/j.rse.2017.06.031, 2017.

Grieger, R., Capon, S. J., Hadwen, W. L., Mackey, B., Grieger, R., Capon, S. J., Hadwen, W. L., and Mackey, B.: Spatial variation and drivers of vegetation structure and composition in coastal freshwater wetlands of subtropical Australia, Marine and Freshwater Research, 72, 1746–1759, https://doi.org/10.1071/MF21023, 2021.

Han, D., Wang, G., Liu, T., Xue, B.-L., Kuczera, G., and Xu, X.: Hydroclimatic response of evapotranspiration partitioning to prolonged droughts in semiarid grassland, Journal of Hydrology, 563, 766–777, https://doi.org/10.1016/j.jhydrol.2018.06.048, 2018.

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N.: The potential of data driven approaches for quantifying hydrological extremes, Advances in Water Resources, 155, 104 017, https://doi.org/10.1016/j.advwatres.2021.104017, 2021.

Hauswirth, S. M., Bierkens, M. F. P., Beijk, V., and Wanders, N.: The suitability of a hybrid framework including data driven approaches for hydrological forecasting, Hydrology and Earth System Sciences Discussions, pp. 1–20, https://doi.org/10.5194/hess-2022-89, 2022.

Hawkins, B. A., Field, R., Cornell, H. V., Currie, D. J., Guégan, J.-F., Kaufman, D. M., Kerr, J. T., Mittelbach, G. G., Oberdorff, T., O'Brien, E. M., Porter, E. E., and Turner, J. R. G.: Energy, water, and broad-scale geographic patterns of species richness, Ecology, 84, 3105–3117, https://doi.org/10.1890/03-8006, 2003.

Holloway-Brown, J., Helmstedt, K. J., and Mengersen, K. L.: Spatial Random Forest (S-RF): A random forest approach for spatially interpolating missing land-cover data with multiple classes, International Journal of Remote Sensing, 42, 3756–3776, https://doi.org/10.1080/01431161.2021.1881183, 2021.

Hoyer, S. and Hamman, J.: xarray: N-D labeled Arrays and Datasets in Python, Journal of Open Research Software, 5, 10, https://doi.org/10.5334/jors.148, 2017.

Huang, S., Tang, L., Hupy, J. P., Wang, Y., and Shao, G.: A commentary review on the use of normalized difference vegetation index (NDVI) in the era of popular remote sensing, Journal of Forestry Research, 32, 1–6, https://doi.org/10.1007/s11676-020-01155-1, 2021.

Jiang, L., Guli Jiapaer, n., Bao, A., Guo, H., and Ndayisaba, F.: Vegetation dynamics and responses to climate change and human activities in Central Asia, The Science of the Total Environment, 599-600, 967–980, https://doi.org/10.1016/j.scitotenv.2017.05.012, 2017.

Jung, M., Dahal, P. R., Butchart, S. H. M., Donald, P. F., De Lamo, X., Lesiv, M., Kapos, V., Rondinini, C., and Visconti, P.: A global map of terrestrial habitat types, Scientific Data, 7, 256, https://doi.org/10.1038/s41597-020-00599-8, bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Biodiversity;Biogeography;Environmental sciences;Macroecology Subject_term_id: biodiversity;biogeography;environmental-sciences;macroecology, 2020.

Li, S., Xu, L., Jing, Y., Yin, H., Li, X., and Guan, X.: High-quality vegetation index product generation: A review of NDVI time series reconstruction techniques, International Journal of Applied Earth Observation and Geoinformation, 105, 102 640, https://doi.org/10.1016/j.jag.2021.102640, 2021a.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions

Li, X., Yuan, W., and Dong, W.: A Machine Learning Method for Predicting Vegetation Indices in China, Remote Sensing, 13, 1147, https://doi.org/10.3390/rs13061147, 2021b.

Liu, Y., Jing, W., Wang, Q., and Xia, X.: Generating high-resolution daily soil moisture by using spatial downscaling techniques: a comparison of six machine learning algorithms, Advances in Water Resources, 141, 103 601, https://doi.org/10.1016/j.advwatres.2020.103601, 2020.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I.: From local explanations to global understanding with explainable AI for trees, Nature Machine Intelligence, 2, 56–67, https://doi.org/10.1038/s42256-019-0138-9, 2020.

McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, in: Proceedings of the 8th Conference on Applied Climatology, vol. 17, pp. 179–183, Boston, issue: 22, 1993.

Meza, I., Siebert, S., Döll, P., Kusche, J., Herbert, C., Eyshi Rezaei, E., Nouri, H., Gerdener, H., Popat, E., Frischen, J., Naumann, G., Vogt, J. V., Walz, Y., Sebesvari, Z., and Hagenlocher, M.: Global-scale drought risk assessment for agricultural systems, Natural Hazards and Earth System Sciences, 20, 695–712, https://doi.org/10.5194/nhess-20-695-2020, 2020.

Moreno-Martínez, , Camps-Valls, G., Kattge, J., Robinson, N., Reichstein, M., van Bodegom, P., Kramer, K., Cornelissen, J. H. C., Reich, P., Bahn, M., Niinemets, , Peñuelas, J., Craine, J. M., Cerabolini, B. E. L., Minden, V., Laughlin, D. C., Sack, L., Allred, B., Baraloto, C., Byun, C., Soudzilovskaia, N. A., and Running, S. W.: A methodology to derive global maps of leaf traits using remote sensing and climate data, Remote Sensing of Environment, 218, 69–88, https://doi.org/10.1016/j.rse.2018.09.006, 2018.

Moussa Kourouma, J., Eze, E., Negash, E., Phiri, D., Vinya, R., Girma, A., and Zenebe, A.: Assessing the spatio-temporal variability of NDVI and VCI as indices of crops productivity in Ethiopia: a remote sensing approach, Geomatics, Natural Hazards and Risk, 12, 2880–2903, https://doi.org/10.1080/19475705.2021.1976849, 2021.

Murray, N. J., Keith, D. A., Bland, L. M., Ferrari, R., Lyons, M. B., Lucas, R., Pettorelli, N., and Nicholson, E.: The role of satellite remote sensing in structured ecosystem risk assessments, Science of The Total Environment, 619-620, 249–257, https://doi.org/10.1016/j.scitotenv.2017.11.034, 2018.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., and Thépaut, J.-N.: ERA5-Land: a state-of-the-art global reanalysis dataset for land applications, Earth System Science Data, 13, 4349–4383, https://doi.org/10.5194/essd-13-4349-2021, publisher: Copernicus GmbH, 2021.

Naumann, G., Barbosa, P., Garrote, L., Iglesias, A., and Vogt, J.: Exploring drought vulnerability in Africa: an indicator based analysis to be used in early warning systems, Hydrology and Earth System Sciences, 18, 1591–1604, https://doi.org/10.5194/hess-18-1591-2014, 2014.

Nguyen, T. T., Ngo, H. H., Guo, W., Chang, S. W., Nguyen, D. D., Nguyen, C. T., Zhang, J., Liang, S., Bui, X. T., and Hoang, N. B.: A low-cost approach for soil moisture prediction using multi-sensor data and machine learning algorithm, Science of The Total Environment, 833, 155 066, https://doi.org/10.1016/j.scitotenv.2022.155066, 2022.

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, : Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, http://jmlr.org/papers/v12/pedregosa11a.html, 2011.

Pretzsch, H., Biber, P., Uhl, E., Dahlhausen, J., Schütze, G., Perkins, D., Rötzer, T., Caldentey, J., Koike, T., Con, T. v., Chavanne, A., Toit, B. d., Foster, K., and Lefer, B.: Climate change accelerates growth of urban trees in metropolises worldwide, Scientific Reports, 7, 15 403, https://doi.org/10.1038/s41598-017-14831-w, 2017.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., and Prabhat: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, https://doi.org/10.1038/s41586-019-0912-1, 2019.

Ribeiro, R. P. and Moniz, N.: Imbalanced regression and extreme value prediction, Machine Learning, 109, 1803–1835, https://doi.org/10.1007/s10994-020-05900-9, 2020.

Rossum, G. v. and Drake, F. L.: The Python language reference, no. Pt. 2 in Python documentation manual / Guido van Rossum; Fred L. Drake [ed.], Python Software Foundation, Hampton, NH, release 3.0.1 [repr.] edn., 2010.

Roy, B.: Optimum machine learning algorithm selection for forecasting vegetation indices: MODIS NDVI & EVI, Remote Sensing Applications: Society and Environment, 23, 100 582, https://doi.org/10.1016/j.rsase.2021.100582, 2021.

Schwalm, C. R., Anderegg, W. R. L., Michalak, A. M., Fisher, J. B., Biondi, F., Koch, G., Litvak, M., Ogle, K., Shaw, J. D., Wolf, A., Huntzinger, D. N., Schaefer, K., Cook, R., Wei, Y., Fang, Y., Hayes, D., Huang, M., Jain, A., and Tian, H.: Global patterns of drought recovery, Nature, 548, 202–205, https://doi.org/10.1038/nature23021, 2017.

Shamshirband, S., Hashemi, S., Salimi, H., Samadianfard, S., Asadi, E., Shadkani, S., Kargar, K., Mosavi, A., Nabipour, N., and Chau, K.-W.: Predicting Standardized Streamflow index for hydrological drought using machine learning models, Engineering Applications of Computational Fluid Mechanics, 14, 339–350, https://doi.org/10.1080/19942060.2020.1715844, 2020.

Sharifi, A.: Yield prediction with machine learning algorithms and satellite images, Journal of the Science of Food and Agriculture, 101, 891–896, https://doi.org/10.1002/jsfa.10696, 2021.

Shen, R., Huang, A., Li, B., and Guo, J.: Construction of a drought monitoring model using deep learning based on multi-source remote sensing data, International Journal of Applied Earth Observation and Geoinformation, 79, 48–57, https://doi.org/10.1016/j.jag.2019.03.006, 2019.

Sicard, P., Agathokleous, E., Araminiene, V., Carrari, E., Hoshika, Y., De Marco, A., and Paoletti, E.: Should we see urban trees as effective solutions to reduce increasing ozone levels in cities?, Environmental Pollution, 243, 163–176, https://doi.org/10.1016/j.envpol.2018.08.049, 2018a.

Sicard, P., Agathokleous, E., Araminiene, V., Carrari, E., Hoshika, Y., De Marco, A., and Paoletti, E.: Should we see urban trees as effective solutions to reduce increasing ozone levels in cities?, Environmental Pollution, 243, 163–176, https://doi.org/10.1016/j.envpol.2018.08.049, 2018b.

Singer, M. B., Asfaw, D. T., Rosolem, R., Cuthbert, M. O., Miralles, D. G., MacLeod, D., Quichimbo, E. A., and Michaelides, K.: Hourly potential evapotranspiration at 0.1° resolution for the global land surface from 1981-present, Scientific Data, 8, 224, https://doi.org/10.1038/s41597-021-01003-9, bandiera_abtest: a Cc_license_type: cc_publicdomain Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Hydrology Subject_term_id: hydrology, 2021.

Smith, N. E., Kooijmans, L. M. J., Koren, G., van Schaik, E., van der Woude, A. M., Wanders, N., Ramonet, M., Xueref-Remy, I., Siebicke, L., Manca, G., Brümmer, C., Baker, I. T., Haynes, K. D., Luijkx, I. T., and Peters, W.: Spring enhancement and summer reduction in carbon uptake during the 2018 drought in northwestern Europe, Philosophical Transactions of the Royal Society B: Biological Sciences, 375, 20190 509, https://doi.org/10.1098/rstb.2019.0509, 2020.

Staben, G., Lucieer, A., and Scarth, P.: Modelling LiDAR derived tree canopy height from Landsat TM, ETM+ and OLI satellite imagery—A machine learning approach, International Journal of Applied Earth Observation and Geoinformation, 73, 666–681, https://doi.org/10.1016/j.jag.2018.08.013, 2018.

Sutanto, S. J., van der Weert, M., Wanders, N., Blauhut, V., and Van Lanen, H. A. J.: Moving from drought hazard to impact forecasts, Nature Communications, 10, 4945, https://doi.org/10.1038/s41467-019-12840-z, 2019.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Tang, L., Chen, X., Cai, X., and Li, J.: Disentangling the roles of land-use-related drivers on vegetation greenness across China, Environmental Research Letters, 16, 124 033, https://doi.org/10.1088/1748-9326/ac37d2, 2021.

Tufaner, F. and Özbeyaz, A.: Estimation and easy calculation of the Palmer Drought Severity Index from the meteorological data by using the advanced machine learning algorithms, Environmental Monitoring and Assessment, 192, 576, https://doi.org/10.1007/s10661-020-08539-0, 2020.

Vereinte Nationen, ed.: Special report on drought 2021, no. 2021 in Global assessment report on disaster risk reduction, United Nations Office for Desaster Risk Reduction, Geneva, 2021.

Vicente-Serrano, S. M., Beguería, S., and López-Moreno, J. I.: A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index, Journal of climate, 23, 1696–1718, 2010.

Vogt, J. V., Naumann, G., Masante, D., Spinoni, J., Cammalleri, C., Erian, W., Pischke, F., Pulwarty, R., and Barbosa, P.: Drought risk assessment and management: a conceptual framework., Publications Office, LU, https://data.europa.eu/doi/10.2760/057223, 2018.

Wanders, N. and Wada, Y.: Human and climate impacts on the 21st century hydrological drought, Journal of Hydrology, 526, 208–220, https://doi.org/10.1016/j.jhydrol.2014.10.047, 2015.

Wang, H., Seaborn, T., Wang, Z., Caudill, C. C., and Link, T. E.: Modeling tree canopy height using machine learning over mixed vegetation landscapes, International Journal of Applied Earth Observation and Geoinformation, 101, 102 353, https://doi.org/10.1016/j.jag.2021.102353, 2021.

Wang, Q., Wang, L., Zhu, X., Ge, Y., Tong, X., and Atkinson, P. M.: Remote sensing image gap filling based on spatial-spectral random forests, Science of Remote Sensing, 5, 100 048, https://doi.org/10.1016/j.srs.2022.100048, 2022.

Xu, Z., Zhou, G., and Shimizu, H.: Plant responses to drought and rewatering, Plant Signaling & Behavior, 5, 649–654, https://doi.org/10.4161/psb.5.6.11398, 2010.

Yamazaki, D., Ikeshima, Daiki, Sosa, Jeison, Allen, George H., Bates, Paul D., and Pavelsky, Tamlin M.: MERIT Hydro: A High-Resolution Global Hydrography Map Based on Latest Topography Dataset, Water Resources Research, 55, 5053–5073, 2019.

Zeng, C., Shen, H., and Zhang, L.: Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method, Remote Sensing of Environment, 131, 182–194, https://doi.org/10.1016/j.rse.2012.12.012, 2013.

Zhang, J., Liu, K., and Wang, M.: Downscaling Groundwater Storage Data in China to a 1-km Resolution Using Machine Learning Methods, Remote Sensing, 13, 523, https://doi.org/10.3390/rs13030523, 2021a.

Zhang, X., Friedl, M. A., Schaaf, C. B., and Strahler, A. H.: Climate controls on vegetation phenological patterns in northern mid- and high latitudes inferred from MODIS data: CLIMATE CONTROLS ON VEGETATION PHENOLOGICAL PATTERNS, Global Change Biology, 10, 1133–1145, https://doi.org/10.1111/j.1529-8817.2003.00784.x, 2004.

Zhang, Y., Keenan, T. F., and Zhou, S.: Exacerbated drought impacts on global ecosystems due to structural overshoot, Nature Ecology & Evolution, 5, 1490–1498, https://doi.org/10.1038/s41559-021-01551-8, 2021b.

Zhao, W. and Duan, S.-B.: Reconstruction of daytime land surface temperatures under cloud-covered conditions using integrated MODIS/Terra land products and MSG geostationary satellite data, Remote Sensing of Environment, 247, 111 931, https://doi.org/10.1016/j.rse.2020.111931, 2020.

Zhu, S., Clement, R., McCalmont, J., Davies, C. A., and Hill, T.: Stable gap-filling for longer eddy covariance data gaps: A globally validated machine-learning approach for carbon dioxide, water, and energy fluxes, Agricultural and Forest Meteorology, 314, 108 777, https://doi.org/10.1016/j.agrformet.2021.108777, 2022.

https://doi.org/10.5194/hess-2022-430
Preprint. Discussion started: 8 February 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Zhu, Z.: Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications, ISPRS Journal of Photogrammetry and Remote Sensing, 130, 370–384, https://doi.org/10.1016/j.isprsjprs.2017.06.013, 2017.