



# Skill of large-scale seasonal drought impact forecasts

Samuel J. Sutanto<sup>1</sup>, Melati van der Weert<sup>1</sup>, Veit Blauhut<sup>2</sup>, and Henny A. J. Van Lanen<sup>1</sup>

<sup>1</sup>Hydrology and Quantitative Water Management Group, Environmental Sciences Department, Wageningen University and Research, Droevendaalsesteeg 3a, 6708 PB Wageningen, the Netherlands

<sup>2</sup>Hydrological Environmental Systems, University of Freiburg, Fahrenbergplatz, 79098 Freiburg, Germany

**Correspondence:** Samuel J. Sutanto (samuel.sutanto@wur.nl)

Received: 2 March 2020 – Discussion started: 9 March 2020

Accepted: 11 May 2020 – Published: 4 June 2020

**Abstract.** Forecasting of drought impacts is still lacking in drought early-warning systems (DEWSs), which presently do not go beyond hazard forecasting. Therefore, we developed drought impact functions using machine learning approaches (logistic regression and random forest) to predict drought impacts with lead times up to 7 months ahead. The observed and forecasted hydrometeorological drought hazards – such as the standardized precipitation index (SPI), standardized precipitation evaporation index (SPEI), and standardized runoff index (SRI) – were obtained from the The EU-funded Enhancing Emergency Management and Response to Extreme Weather and Climate Events (ANYWHERE) DEWS. Reported drought impact data, taken from the European Drought Impact Report Inventory (EDII), were used to develop and validate drought impact functions. The skill of the drought impact functions in forecasting drought impacts was evaluated using the Brier skill score and relative operating characteristic metrics for five cases representing different spatial aggregation and lumping of impacted sectors. Results show that hydrological drought hazard represented by SRI has higher skill than meteorological drought represented by SPI and SPEI. For German regions, impact functions developed using random forests indicate a higher discriminative ability to forecast drought impacts than logistic regression. Moreover, skill is higher for cases with higher spatial resolution and less lumped impacted sectors (cases 4 and 5), with considerable skill up to 3–4 months ahead. The forecasting skill of drought impacts using machine learning greatly depends on the availability of impact data. This study demonstrates that the drought impact functions could not be developed for certain regions and impacted sectors, owing to the lack of reported impacts.

## 1 Introduction

Drought is a creeping phenomenon that slowly covers and affects extensive areas (Wilhite, 2000; Tallaksen and Van Lanen, 2004; Mishra and Singh, 2010; Van Loon, 2015). It can last from months to years and can have multi-faceted impacts on, for example, crop production, drinking water supply, electricity production, waterborne transportation, and recreation (Wilhite, 2000; UNDP, 2006; EEA, 2007; Stahl et al., 2016). In terms of the number of affected people and economic losses, drought is categorized as one of the most damaging natural hazards (Wilhite and Vanyarkho, 2000; EEA, 2012). Worldwide, 2 billion people were affected and more than 10 million people passed away due to the impacts of drought between 1900 and 2010 (Georgi et al., 2012). In the period 1976–2006, 15 % of the European territory and 17 % of the European population was affected by drought on an average annual basis, and it was estimated that the total damage amounted to > EUR 100 billion (EEA, 2007).

Actions – such as the development of drought monitoring and early-warning systems, as it is highlighted in the UN Sendai Framework for Disaster Risk Reduction (Poljanšek et al., 2017) – have been taken to reduce the impacts of drought. However, present drought monitoring and early-warning systems only provide the end users with information of present and forecasted drought hazard conditions, mainly derived from the hydro-meteorological drought indices (Sepulcre-Canto et al., 2012; Pozzi et al., 2013; Sheffield et al., 2014; Sutanto et al., 2019a). Information on forecasted drought conditions, expressed by a drought hazard index, does not directly translate the hazard into impacts. Thus, the forecasting of drought impacts is still missing in drought early-warning systems. This may be due to the fact

that finding information on drought impacts is quite challenging, as drought occurs in different hydrological domains and has mainly non-structural consequences (Lackstrom et al., 2013; Stahl et al., 2016; Kreibich et al., 2019). Nevertheless, the likelihood of impact occurrence – which is a result of the hazard, exposure, and vulnerability – is essential to stakeholders in order to manage drought events. Relating drought hazard indices to their impacts is beneficial for decision making and risk management (Hayes et al., 2011). For water managers, stakeholders, and policymakers, there is a vital need to be informed about forecasted impacts of drought events, rather than only about the drought hazard. Thus, adaptation actions can be prepared and taken well in advance, to reduce the impact of drought.

As an alternative to a conventional bottom-up approach that links potential risks/impacts to hazard, exposure, and vulnerability (e.g., Carrão et al., 2016; Naumann et al., 2014, 2015), some studies have been conducted to develop a link between drought impacts and the underlying hazard (top-down approach). Gudmundsson et al. (2014) were among the first to study this relationship using the standardized precipitation index (SPI; McKee et al., 1993) and the European Fire Database (EDF; Camia et al., 2014). This study in southern Europe showed that there is a significant relationship between the probability of above-normal wildfire incidents and meteorological drought hazards. The development of the European Drought Impact Report Inventory (EDII; Stahl et al., 2016) – a unique collection of categorized, temporally, and spatially assigned drought impact information – offers the novel possibility to evaluate drought impacts. Using the EDII database, Blauhut et al. (2015) assessed the link between drought hazard and various impacted sectors across Europe on an annual scale, such as agriculture and livestock farming, energy and industry, public water supply, and water quality. Stagge et al. (2015) enhanced this link by accounting for seasonality, incorporating interannual trends, and allowing for nonlinear effects. In a joint publication, Blauhut et al. (2016) developed a method for testing currently monitored drought hazard indices and vulnerability information on their ability to model the likelihood of impact occurrence. All the studies described above used similar methods to derive predictor–impact relations, which were based on the machine learning method logistic regression (LR). In addition, Bachmair et al. (2017) assessed further machine learning approaches to deriving the link between drought hazard and their impacts, such as zero-altered negative binomial regression and random forest (RF) models.

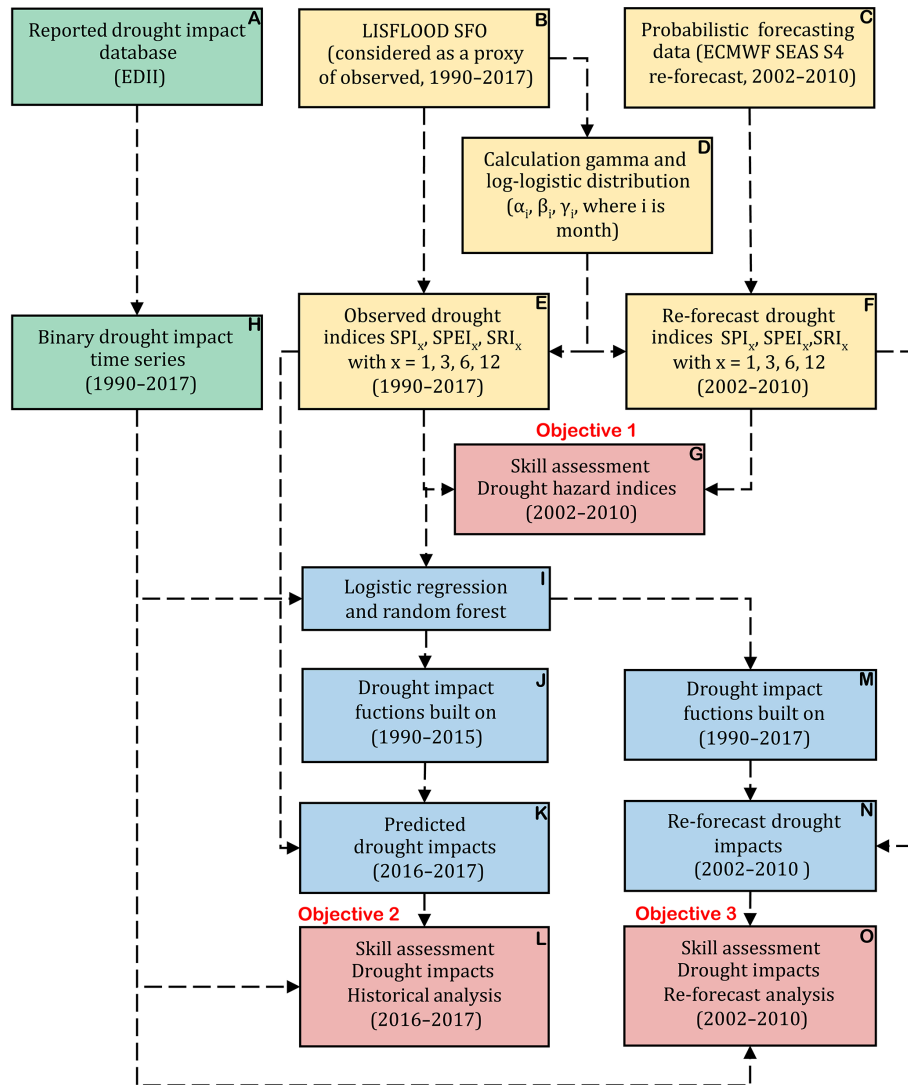
The aforementioned studies in general used historical observational data to derive the meteorological drought indices (SPI and standardized precipitation evaporation index, SPEI; Vicente-Serrano et al., 2010) to construct drought impact functions. The results were promising, and hence it was hypothesized that these also could be implemented in the forecast mode. Sutanto et al. (2019b), therefore, have conducted a first study that explored the possibility of forecasting the

drought impacts using hydro-meteorological drought indices – including the SPI, SPEI, and standardized runoff index (SRI, Shukla and Wood, 2008) – and the EDII database with data from Germany. They only used the RF method to forecast drought impacts for four different lumped impacted sectors (i.e., impact groups). Their study shows that, in general, forecasting drought impacts is feasible, with skill up to a few months ahead depending on the number of reported impacts and drought duration. However, the skill likely depends on the machine learning approach adopted to derive functions that relate drought hazard indices to impacts (i.e., impact functions).

In this study, we enhance the previous studies by evaluating the skill of seasonal drought impact forecasts using the RF and LR approaches for different combinations of spatial aggregations and impact categories (i.e., cases), which is the novelty of our paper. The RF and LR approaches have proven value in developing drought impact functions using the historical data and have not been used yet to study drought impact forecasting using dynamical weather forecasts. One should note that a study conducted by Sutanto et al. (2019b) only used the RF approach and lumped impacted sectors. To evaluate the skill of seasonal drought impact forecasts using the RF and LR approaches, we divided the study into three parts with the following objectives (Fig. 1): (1) evaluation of the skill in forecasting drought hazards using re-forecast data from 2002 to 2010 (box G), which is a first step in the forecasting chain; (2) evaluation of the skill of the drought impact functions based upon the RF and LR approaches using proxy observed data and a split-sampling method (box L); and (3) testing the skill of developed drought impact functions based upon the RF and LR approaches to forecasting drought impacts using re-forecast data from 2002 to 2010 (box O).

## 2 Data and methods

The flowchart that summarizes this section, as well as the objectives of the three parts of the study, is presented in Fig. 1. In general, two datasets and two machine learning approaches were used in this study. The hydro-meteorological data (yellow boxes in Fig. 1) were required to calculate the drought hazard indices (SPI-x, SPEI-x, and SRI-x). The drought impact data (green box in Fig. 1) were a prerequisite to translating the drought hazard into the likelihood of drought impact occurrence (LIO) in the next phase using the binary LR model and the RF model (blue boxes in Fig. 1). Three forecast skill analyses were carried out to assess (1) the drought hazard forecast skill (objective 1), (2) impact forecasting skill by using observed impact data and a split-sampling technique (objective 2), and (3) drought impact forecasting skill by using re-forecast data (objective 3).



**Figure 1.** Workflow diagram, including colored boxes to illustrate the main components (drought hazard: yellow; drought impact: green; linking impact to hazard: blue; and validation of forecasting skill: red), as well as the objectives of the three parts of the study.

## 2.1 The EDII database

The EDII is a joint effort of the EU FP-7 project DROUGHT-R&SPI (<http://europeandroughtcentre.com/>, last access: June 2020), which is actually hosted, updated, and maintained by the University of Freiburg (Stahl et al., 2016). The database is a compilation of drought impact reports in Europe, which stem from a variety of sources, for example governmental reports, scientific publications, media, and questionnaires (box A in Fig. 1). In the EDII, drought impacts are considered as negative environmental, economic, or social effects of the hazard of drought. Hence, drought conditions – such as an anomaly in precipitation, soil moisture, groundwater levels, or runoff – which do not result in a negative consequence are not considered as an impact in the EDII inventory. Each impact report is assigned to (1) temporal ref-

erence (this can be indicated by month, season, or year as a minimum), (2) spatial reference (the location of the reported impacts can be either referred to by different levels of geographical regions using the European Union Nomenclature of Territorial Units for Statistics (NUTS) (Eurostat, 2011) or specified by rivers or lakes), and (3) one of 15 impact categories and a selection of 105 impact types (see Stahl et al. (2016) for detailed information on the EDII database). In this study, drought impact reports from the period 1990–2017 for Germany at the country and county level (NUTS-1 regions, Fig. S1 in the Supplement for the Germany NUTS-1 regions) were used. For each NUTS region and impacted category, binary time series (0: no drought; 1: drought) of monthly reported drought impacts were derived from the EDII (1990–2017, box H in Fig. 1) following Blauhut et al. (2015) and Bachmair et al. (2017).

Since the availability of impact information from the EDII database appears to vary across Europe (Stahl et al., 2016; Sutanto et al., 2019b), we developed drought impact functions for the five different cases, i.e., combinations of spatial aggregations and impact categories. For cases 1 and 2 we accumulated the entire country of Germany, and for cases 3–5, i.e., NUTS-1 regions (Fig. S1), a higher spatial resolution was applied. For cases 2 and 5, all 15 individual impacted sectors were tested to forecast drought impacts. For cases 1 and 3 all sectors were lumped. For case 4, the EDII impacted sectors were arbitrarily grouped as follows: group 1 consists of agriculture and livestock farming and forestry; group 2 consists of energy and industry, transportation, and public water supply; group 3 consists of water quality, freshwater ecosystems, and terrestrial ecosystems; and group 4 consists of wildfire, air quality, and human health and public safety. We grouped the impacts based on their likely relation to specific drought types (meteorological, agricultural, and hydrological drought) and topic (water quality and ecosystems). Group 1 relates to agricultural drought, group 2 relates to streamflow drought (hydrology), group 3 relates to hydrological drought linked to ecosystems, and group 4 relates to meteorological drought and heat. These cases and the selection of the EDII impacts are documented in Table 1. Hence, the binary time series differ and are case specific (the more impact categories and regions merged, the more reported impacts reported). The impacts that were reported as a season were transformed into months of that season (winter: December, January, and February; spring: March, April, and May; and so on).

## 2.2 Hydro-meteorological dataset

The hydro-meteorological data used in this study were obtained from the European Flood Awareness System (EFAS; Thielen et al., 2009; Pappenberger et al., 2011) with a spatial resolution of 5 km by 5 km. The EFAS used in the study is based upon a state-of-the-art LISFLOOD rainfall–runoff model (Bartholmes et al., 2008; Van Der Knijff et al., 2010; Burek et al., 2013) and the European Centre for Medium-Range Weather Forecasts (ECMWF) seasonal forecasts S4 (SEAS4; Smith et al., 2016). Gridded observed meteorological data, such as precipitation and evapotranspiration, were used by the LISFLOOD model to simulate gridded runoff. In EFAS, potential evapotranspiration is calculated through the offline LISVAP pre-processor based on the Penman–Monteith equation (Van Der Knijff, 2008). The gridded meteorological and runoff data (simulation forced with observations, SFO; hereafter referred to as proxy observed data; box B in Fig. 1) were used to calculate the observed drought hazards on a timescale of 1990–2017 (box E in Fig. 1). Simulated runoff had to be used since gridded runoff cannot be measured. Streamflow observational data in some rivers are available (> 250 catchments across Europe), and these were used to calibrate and validate the LISFLOOD hydrological

model. LISFLOOD obtained a median Nash–Sutcliffe efficiency (NSE) of 0.57 over the validation period (Feyen and Dankers, 2009; Forzieri et al., 2014).

Meteorological and hydrological re-forecasts (in many cases referred to as hindcasts) were obtained from the ECMWF meteorological forecast SEAS4 and the LISFLOOD model forced with the ECMWF SEAS4, respectively (box C in Fig. 1). These re-forecast data were used to calculate the forecasted drought hazards from 2002 to 2010 (box F in Fig. 1), using the distribution parameters derived from the observed datasets (box D in Fig. 1). Re-forecasts of meteorological and hydrological variables have a lead time (LT) of 215 d (circa 7 months) and consist of 15 ensemble members. Detailed information on the hydro-meteorological dataset from EFAS can be found in Arnal et al. (2018), Van Hateren et al. (2019), and Sutanto et al. (2019a). The daily proxy observed and re-forecast data were transferred into monthly time series.

## 2.3 Standardized drought hazard indices

The calculation of hydro-meteorological drought hazards was carried out using the SPI (McKee et al., 1993), the SPEI (Vicente-Serrano et al., 2010), and the SRI (Shukla and Wood, 2008), with accumulation periods of 1, 3, 6, and 12 months (boxes E and F in Fig. 1). The standardized indices were calculated by fitting a probabilistic distribution on monthly observed variables (1990–2017): precipitation for SPI, precipitation minus potential evapotranspiration for SPEI, and runoff for SRI. These input data were averaged for the German counties (NUTS-1) and country before the calculation of the standardized indices was performed. A gamma distribution (SPI and SRI) and three-parameter log-logistic distribution (SPEI) were selected to transform the monthly hydro-meteorological data into 144 distributions for every month of the year, for each drought hazard index, and for each accumulation period ( $12 \times 3 \times 4$ , box D in Fig. 1). We used the distributions to calculate monthly time series of the three drought hazard indices (SPI, SPEI, and SRI) for the observed data (1990–2017) (box E in Fig. 1) and re-forecast data (2002–2010, box F in Fig. 1).

## 2.4 Machine learning for modeling drought impacts

LR predicts the probability of a dichotomous variable. This model belongs to the class of generalized linear models (GLMs; Zuur et al., 2009). The LIO (logit-transformed) is modeled as a linear function of drought hazard predictors as

$$\log \frac{\text{LIO}}{1 - \text{LIO}} = \alpha + \sum_i \beta_i x_i. \quad (1)$$

The left side of the equation represents the logit transformation of LIO. The model parameter  $\alpha$  and  $\beta$  are intercept and coefficient, respectively, that were estimated using standard regression techniques within the framework of GLMs (Peng

**Table 1.** Different cases for which drought impacts were forecasted. For the German NUTS-1 regions, see Fig. S1.

Case	Spatial resolution	Impact	Ideal number of impact functions per method	Realization impact number of functions per method
1	Germany	Impacts of all impacted sectors lumped	$1 \times 1 = 1$	1
2	Germany	Individual impact category	$1 \times 15 = 15$	9
3	NUTS-1 level	Impacts of all impacted sectors lumped	$16 \times 1 = 16$	12
4	NUTS-1 level	Impacted sectors lumped in four groups	$16 \times 4 = 64$	19
5	NUTS-1 level	Individual impact category	$16 \times 15 = 240$	15

et al., 2002).  $x$  is predictor variables (e.g., month, year, SPI- $x$ , SPEI- $x$ , and SRI- $x$ ) with different accumulation periods, and  $i$  is the number of predictor variables. The months of impact occurrence (January, February, and so on until December) were included to account for seasonality, as some drought impacts are connected to particular seasons. The years of impact occurrence were included (e.g., 2003, 2010, 2015, and so on) as a predictor since drought impacts gradually have been better documented throughout the years. To find the best LR models for LIO (boxes J and M in Fig. 1), we applied the following three steps following Blauhut et al. (2016). (1) A selection from the initial predictors (SPI- $x$ , SPEI- $x$ , SRI- $x$ , month, and year) was tested for their single ability to predict the impact. Insignificant predictors indicated with a probability value ( $p$ ) larger than 0.05 were omitted, and the remaining predictors were tested for collinearity (step 2). (2) Each combination of two independent variables was tested for their collinearity. For predictors which were highly correlated to each other ( $p > 0.7$ ), the best was kept. Thereafter, (3) the forward and backward stepwise selection was used to find the best combination of predictors defined by the Akaike information criterion (AIC). This analysis was performed using the “MASS” package in RStudio (Ripley et al., 2019). In this procedure, the model started with no independent variables. Sequentially, the most significant independent variable was added to the model and was examined for the model improvement. The independent variable was removed from the model if it did not increase the model performance. This procedure was performed until all selected independent variables from steps 1 and 2 were analyzed. The selection procedures of the best predictors mean that not all predictors are used in the LR method. To estimate the contribution of each variable to the model, the absolute value of the  $t$  statistic for each model parameter was assessed using the “caret” package in RStudio (Kuhn, 2019).

RF is a machine learning algorithm that is powerful for developing a predictive model (Breiman, 2001). The system creates a multitude of random independent decision trees as an ensemble to reduce the chance of overfitting and sensitivity to training data configuration (in our case 2000 trees). Each tree is constructed based on a bootstrapped fixed size

of a subsample of the data, and predictions are made by the mean prediction of the individual trees. The RF method was performed using the “randomForest” package in RStudio (Liaw and Wiener, 2002). We used all the predictor variables in the RF model (box E in Fig. 1), and the response variables consist of the binary time series of reported drought impacts from the EDII (box H in Fig. 1). We did not cross-validate our model; instead, we used the caret feature to identify the drought hazard indices best linked to impact occurrence (Kuhn, 2019). Caret uses the prediction accuracy on the out-of-bag (OOB) portion for both the full model and after permuting each predictor variable. The differences between the two models are then averaged over all trees and normalized by the standard error. By doing this, the performance of the model was tested. LIO was estimated by calculating the probability of the number of trees that indicated the impact for every German NUTS region. We trained the machine learning models (LR and RF) using the data from 1990 to 2015 for objective 2 (box J in Fig. 1). For objective 3, we used all observations from 1990 to 2017 for model training (box M in Fig. 1). Bachmair et al. (2016, 2017) and Sutanto et al. (2019b) provide a detailed explanation of the use of RF to develop drought impact forecasting functions.

## 2.5 Forecasting skill score

The skill of forecasted drought hazard indices and drought impacts was evaluated using common skill score metrics, such as the Brier skill score (BSS; Brier, 1950) and relative operating characteristic (ROC; Mason, 1982). The BSS can be used to measure both the reliability and sharpness of forecasts (Trambauer et al., 2015; Arnal et al., 2018). For the BSS, both the Brier score for the climatology (BSclim) and the Brier score for the forecast (BSf) are required. The Brier score (BS) measures the mean squared difference between the predicted probability and the verification sample. For BSclim, a threshold of  $-0.5$  was chosen to give a good balance between capturing either too many minor droughts or too few drought events (Trambauer et al., 2015). The standardized index of  $-0.5$  stems from a normal distribution, and therefore the value for  $p$  is 0.3085 for every month

(Van Hateren et al., 2019). The BSS measures the improvement of the probabilistic forecasts relative to climatology. It has a range of  $[1, -\infty]$ , where 1 indicates the highest skill and values below 0 indicate that the probabilistic forecast does not show additional predictive performance against the climatology.

The ROC curve was used as a criterion to measure the discriminate ability between two outcomes (Mason and Graham, 1999). The ROC curve gives the relation between the true positive rate (sensitivity) and the false positive rate (specificity). The area under the curve (AUC) was calculated to measure the accuracy of the forecast. The AUC has a range from 0 to 1, with a perfect score of 1.  $AUC = 0.5$  is as good as flipping a coin, and  $AUC < 0.5$  is unskillful.

### 3 Results

#### 3.1 Skill of drought hazard forecasts

The skill of probabilistic drought hazard forecasts (SPI- $x$ , SPEI- $x$ , and SRI- $x$ ) for our example, Germany, using the median of the ensemble of re-forecast data from 2002 to 2010 is presented in Fig. 2. The skill was measured using the BSS and ROC skill scores and presented for forecasts done in each season (colored lines) and on an annual timescale (black line). Figure 2 shows that drought indices with longer accumulation periods perform better than ones with short accumulation periods. For example, SPI-1 forecasts using the BSS show no skill for all seasons except for forecasts made in autumn for a LT of 1 month. On the other hand, the skill much improves for longer accumulation periods, i.e., SPI-12, with BSS values around 0.7 for 1-month LT. The lowest skill is found for forecasts done in summer for all accumulation periods. As expected, the skill also decreases for higher LTs.

The predictive skill of drought hazard forecasts varies between the drought indices. In general, SPI- $x$  and SPEI- $x$  have similar predictive skill for long accumulation periods (e.g.,  $x = 6$  and 12 months). However, the SPEI- $x$  index shows higher skill than the SPI- $x$  for short accumulation periods. The SPI-3 forecasted during summer shows a BSS value of 0.75, while the SPEI-3 shows a BSS value of 0.85. In general, the SPEI drought forecasts have slightly higher skill than the SPI drought forecasts. The source of predictability of SPEI comes from the higher temperature predictability due to the North Atlantic Oscillation (NAO) in Europe than precipitation (Vitart, 2014). Temperature is one of the weather variables that controls potential evapotranspiration. The hydrological drought index, SRI- $x$ , predicts drought better than the meteorological drought indices, for example SPI- $x$  and SPEI- $x$ . This is illustrated in the BSS and ROC figures. The BSS for SRI-1, for example, shows that there is a predictive value up to a LT of 2 months, which indicates a similar skill to SPI-3 and SPEI-3. For drought indices with longer accumulation periods, SRI- $x$  performs better for almost all LTs

than SPI- $x$  and SPEI- $x$ . The comparison between the BSS and ROC skill score shows that drought forecasts using the ROC method produces a higher score. This is clearly seen for drought forecasts with higher accumulation periods and LTs (Fig. 2).

#### 3.2 Skill of drought impact forecasts using observed data

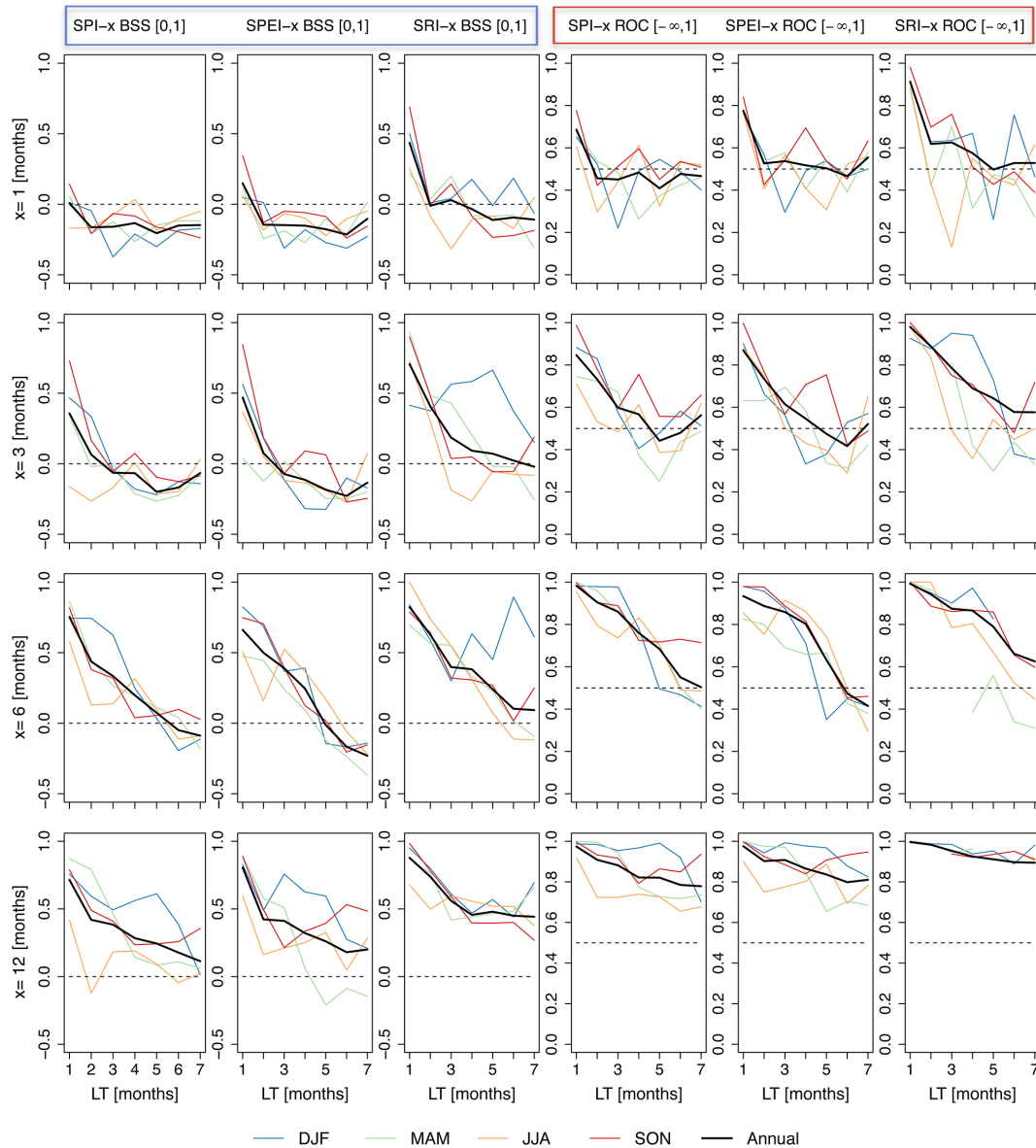
To test drought impact function performance for the five different cases (box L in Fig. 1 and Table 1), we developed these functions using the observed data from 1990 to 2015 (box J in Fig. 1). The predictive power of the model is assessed by comparing it with observed impacts from the last 2 years (2016 to 2017, split-sampling technique). Please note that no re-forecast data were used in this analysis. A summary of the model performance using the ROC skill score for the different cases is presented as box plots in Fig. 3. Please keep in mind that case 1 only covers one impact function, which is developed using all impact data lumped, irrespective of the impacted sector, for Germany (all NUTS-1 regions). Hence, case 1 is presented as a single line in Fig. 3. For each box plot, the median of the ROC values obtained from all models is also presented as a single line.

Figure 3 shows that, overall, the models have skills in distinguishing impact and no impact. This is indicated by the median for all five cases above 0.5. The analysis of the medians between the two machine learning approaches shows a large difference in skill score for case 1. The model's skill score for LR ( $ROC = 0.53$ ) is substantially lower than RF ( $ROC = 0.7$ ). However, the two methods show similar skills for the other cases, which have a substantially lower number of impact reports. Cases 2 and 3 show lower performance than cases 4 and 5. Variation within the cases seems more pronounced in LR than in RF.

#### 3.3 Skill of drought impact forecasts using re-forecast data

After testing the drought impact function performance using a split-sampling technique (Sect. 3.2), we re-developed drought impact functions using all observed datasets from 1990 to 2017 (box M in Fig. 1). These models were built as if these would be implemented in a drought early-warning system. Another reason to use the full dataset is that we would like to train the models with as many data as possible. After developing the functions, these were used to forecast the drought impacts (box N in Fig. 1) using the re-forecast data from 2002 to 2010 (box F in Fig. 1). A summary of the skill of drought impact functions in forecasting drought impacts (box O in Fig. 1) using the ROC skill score is presented in Fig. 4.

In Fig. 4, we divide the analysis of drought impact forecasting skill into short LTs (1–3 months) and long LTs (4–7 months). Obviously, forecasting drought impacts for

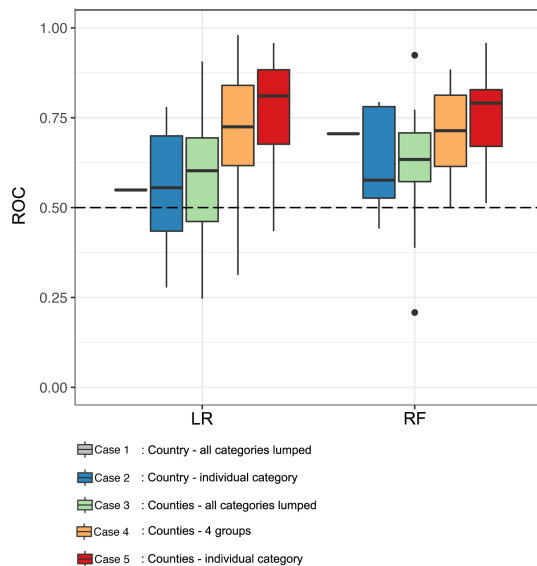


**Figure 2.** Skill scores of re-forecast drought hazard indices (SPI- $x$ , SPEI- $x$ , and SRI- $x$ ) for 7-month lead times (LTs) in Germany for the period 2002–2010 (median of 15 ensemble members). The four different rows of graphs represent four different accumulation periods ( $x = 1, 3, 6,$  and  $12$  months). The black dashed line indicates the threshold (BSS = 0; ROC = 0.5). The first three columns show the BSS for the three different drought indices (indicated by a blue box at the top). The last three columns are for ROC (indicated by a red box at the top).

shorter LTs produce a higher skill than longer LTs. This is true for all cases and the two modeling approaches. Figure 4 also shows that the RF performs better than the LR for all LTs. If only the median value is considered, models developed using the LR method have skill only for cases 3–5 and only for short LTs. For long LTs, all the LR models show no skill. For RF, however, skill is obtained for both short and long LTs and for all cases. Variation within cases seems to be smaller for RF than LR, and the differences are larger for longer LTs than for shorter ones. The skill of drought

impact forecasts using LR for short LTs (cases 3–5) varies around 0.74, while the skill using RF fluctuates around 0.85.

As discussed in the previous paragraph, the highest skill is achieved using the RF method for cases 4 and 5. Therefore, the skill of drought impact forecasts developed using RF for case 4 is presented in Fig. 5, for each German NUTS-1 region and different impact groups. We only present the skill of drought impact forecasts for case 4 because the number of drought impact functions that can be developed for case 5 is limited. This is due to the large differences in the number of reported impacts for each impacted sector between the



**Figure 3.** Model performance for the five cases using logistic regression (LR) and random forest (RF) approaches to predict drought impacts for the period 2016–2017. The boxes indicate the spread of ROC values for drought impact functions (15 ensemble members). The black dashed line indicates the threshold (ROC = 0.5). All values below this line show that the models have no discriminative capacity to distinguish impact and no impact. The dots stand for model outliers.

German NUTS-1 regions (Stahl et al., 2016; Sutanto et al., 2019b). For example, for case 5, drought impacts on agriculture and livestock farming or forestry can be forecasted in the Baden-Württemberg (BW) and Bavaria (BV) regions only, respectively (figure not shown; please see Fig. S1 for German NUTS-1 regions). However, if we combine these two impacted sectors into one group (group 1, case 4), impact functions can be developed for regions BW, BV, Rheinland-Pfalz (RP), and Niedersachsen (LS) (Fig. 5).

Figure 5 shows that, even if we combine several impacted sectors into four different impact groups, drought impact functions can be developed only for some regions in Germany due to the aforementioned data limitation and regional characteristics. For example, Hamburg (HH) and Bremen (HB) are mainly cities; hence, there is no reported agriculture and livestock farming impact. Based on the BSS skill score, good drought impact forecasting skill is achieved only for impact group 2 in BV. However, the ROC skill score shows a good skill for all impact groups for short LTs. In general, the skill of drought impact forecasts using BSS is fair for short LTs and poor for longer LTs, implying that for shorter LTs a reasonable distinction can be made between impact and no impact. This is especially the case for situations where the observed data show prolonged periods with drought impacts (Sutanto et al., 2019b).

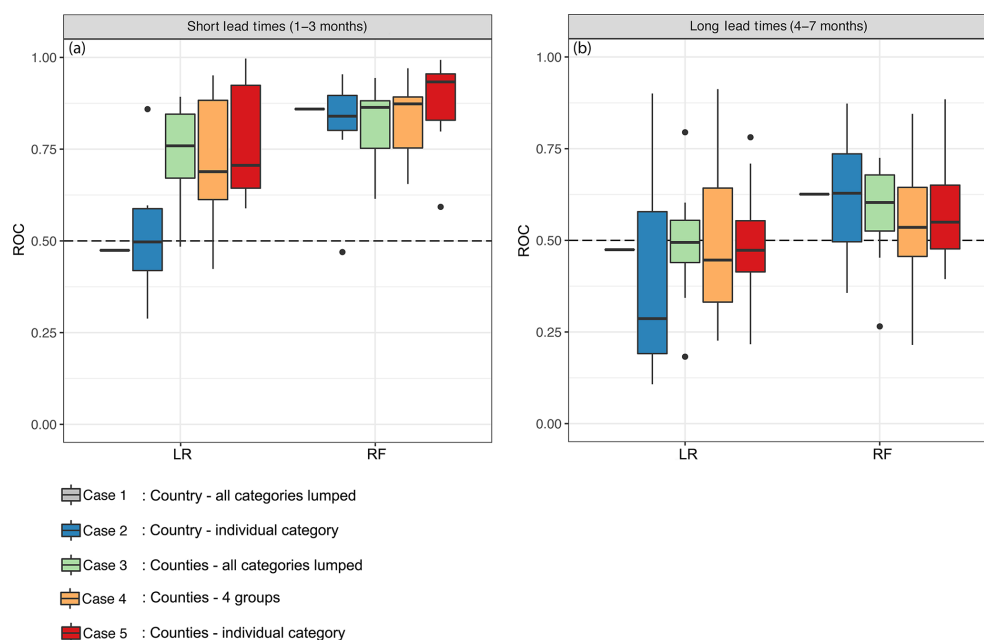
## 4 Discussion

The use of standardized drought indices – i.e., SPI, SPEI, and SRI – requires preferably a 30-year record (McKee et al., 1993; Vicente-Serrano et al., 2010; Shukla and Wood, 2008). In our study, we could only use 28 years of observational record (proxy) from 1990 to 2017 (Sect. 2.2). EFAS data before 1990 are not available, and we only had data up to 2017. The length of the observational data might have some implications on the calculation of parameters of the monthly distributions, which affects the calculation of the drought severity index. For example, if the records (< 30 years) do not include extreme low/high events, then the calculation of drought indices will overestimate these low/high severities due to lack of low/high data in the head/tail of the normal distribution. In our study, 2 years of missing data in the record do not significantly influence our results for the following reasons: (1) the years 1988 and 1989 were not recorded as extreme drought years, and (2) we included drought years in parts of Europe, for example 1991–1992, 1995, 1996–1997, 2003, 2006, 2008, and 2015 (Spinoni et al., 2015).

Analysis of model performance using the split-sampling method (Fig. 3) indicates that the two machine learning approaches (LR and RF) seem to have similar descriptive power. However, for drought impact models evaluated using re-forecast data, RF shows better performance, denoted by higher ROC scores (Fig. 4). For evaluation using re-forecast data, the drought impact models were trained using all observed data from 1990 to 2017. This implies that RF seems to have a longer memory than LR, meaning that RF keeps memory of the trained data, which were used to build the drought impact functions. Moreover, the number of predictors used in RF and LR differs because of the selection processes in the LR method, as done in previous studies (Stagge et al., 2015; Blauhut et al., 2015), whereas in RF we used all predictors (Bachmair et al., 2016).

Figures S2 and S3 provide the importance of predictors in LR and RF, respectively. Figure S2 clearly shows that in the LR method only some of the predictors were used, whereas in the RF method all the predictors were used (Fig. S3). For example, in BV for impact group 2, LR only used SPI-3, SPEI-12, and year as the predictors, with SPEI-12 as the most important predictor followed by SPI-3 and year. Using the RF method to predict drought impact group 2 in BV, the most important predictors are SPEI-6, SRI-6, and year, while the less important predictors are SPI-1, SPI-12, and SPEI-1. In this study, we did not use the same numbers of predictor for both LR and RF; instead, we followed the same approach as in previous studies (Stagge et al., 2015; Blauhut et al., 2015; Bachmair et al., 2016, 2017). Whether the skill of drought impact functions developed using LR (RF) would improve (decrease), if we trained the models using the same number of predictors, is subject to future study.



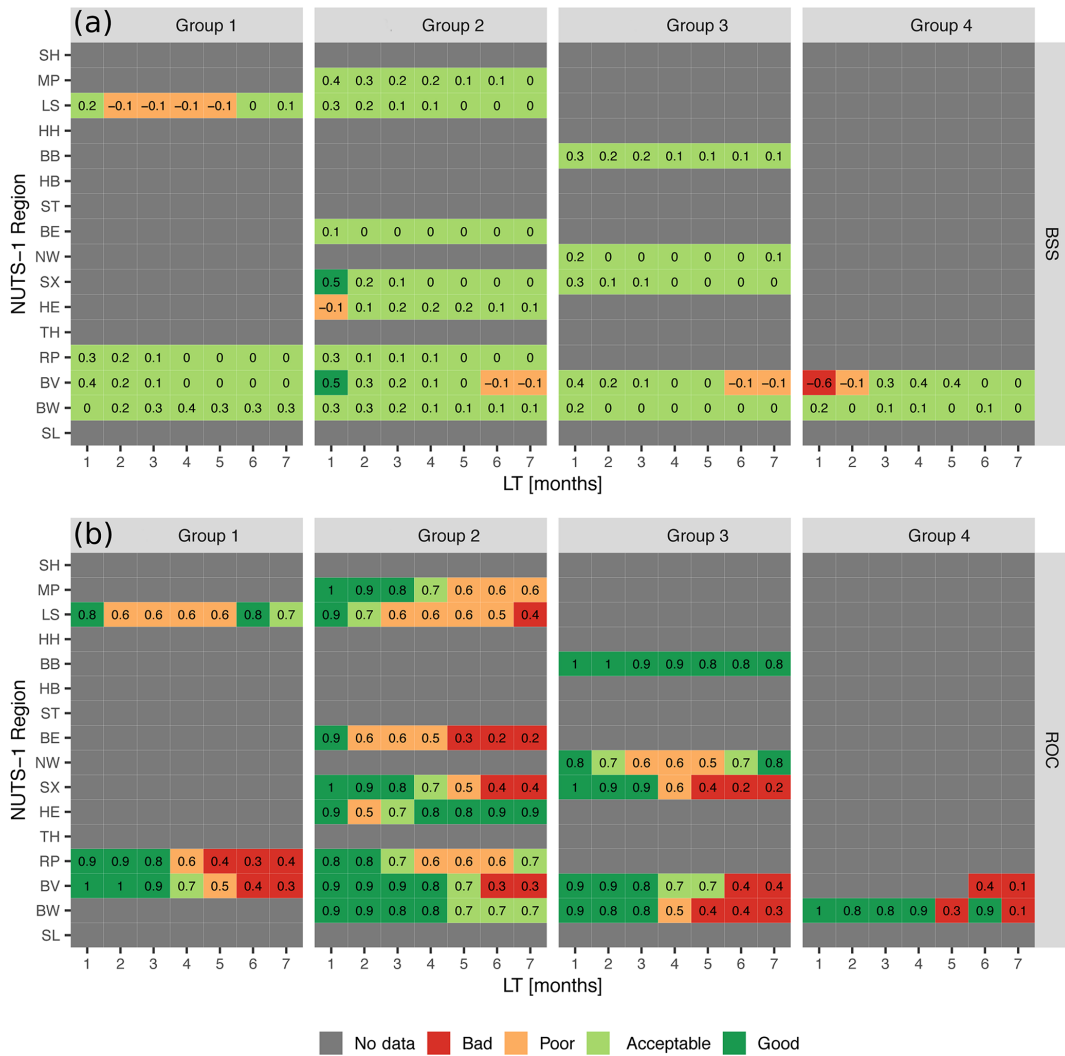


**Figure 4.** Skill of re-forecast drought impacts for the five cases using logistic regression and random forest approaches for the period 2002–2010. The boxes indicate the spread of ROC values for drought impact functions (15 ensemble members) for (a) short lead times (1–3 months) and (b) long lead times (4–7 months). The black dashed line indicates the threshold. All values below this line show that the models have no discriminative capacity to distinguish impact and no impact. The dots stand for models outliers.

It is also worth mentioning that the RF has several advantages relative to the LR in this study: (i) it can handle non-linear relationships among variables, (ii) it is robust to noise, (iii) the bootstrap sampling provides a way to account for the uncertainty in the impact data, and (iv) it has a higher true positive rate (Breiman, 2001; Hastie et al., 2009; Evans et al., 2011; Trigila et al., 2015; Kirasich et al., 2018; Fung et al., 2019). Among some advantages of the RF approach mentioned above, the ability of RF to handle nonlinearity is of the utmost importance in drought study. The drought severities obtained from the standardized indices (i.e., SPI, SPEI, and SRI), i.e., input data, were derived from functions that were developed by fitting the gamma distribution and that later were transformed into normal distribution (Sect. 2.3). This indicates that the drought severities, as input data for RF and LR, are not linear, which is conjectured to be the reason for better skill of RF than LR. Furthermore, the RF has the potential to handle information on the severity of the impact, which is beneficial for future decision making (Breiman, 2001; Bachmair et al., 2016, 2017). Our results support previous drought studies, which concluded that RF performs better than any other machine learning methods, such as LR, boosted regression trees, cubist, decision trees, and hurdle (Park et al., 2015; Bachmair et al., 2017; Rhee and Im, 2017).

Machine learning is a data-driven approach, in which the model performance is strongly related to the quality and quantity of the underlying data (Solomatine and Shrestha, 2009; Elshorbagy et al., 2010; Mount et al., 2016). Drought

impact functions could not be developed in HH, HB, ST, Thüringen (TH), and Saarland (SL) or any case because there are fewer than 20 months with reported impacts in the period 1990 to 2017 (Fig. S4). Many reports were collected in the south of Germany, while only a few reports were collected in the north of Germany. For example, many reports were found in BV and BW (> 120 months) during the study period of 1990–2017. However, fewer than 20 months of reports were collected in SH, HH, HB, ST, BE, TH, and SL (red line Fig. S4). The total number of reported impacts in the whole of Germany from 1990 to 2017 is  $\sim 1160$ . Combining all impact data from all impacted sectors for a NUTS-1 region helps to build a drought impact function in a region that has a limited number of reported impacts. However, if impacts on all sectors are lumped (case 3), then, obviously, the model only provides total impact occurrence (impact or no impact) without any information on impacted sectors. For example, a drought impact model can be built in the SH region, which has a total number of reported impacts of around 20 coming from eight different impacted sectors. In this case, the model can predict the impact occurrence relatively well, with a ROC value of 0.8 for  $LT = 1$  (figure not shown). When we break down all impacts for a NUTS-1 region into impact groups (case 4), meaning fewer reported impacts for each group, a drought impact model cannot be made in this region (Fig. 5). Thus, a good balance between the number of groups and reported impacts in each group by considering similar



**Figure 5.** Skill of re-forecast drought impacts for four impact groups and each German NUTS-1 region using the random forest approach for the period 2002–2010. The performance is measured using (a) BSS and (b) ROC for lead times up to 7 months. For the acronym of NUTS-1 regions, see Fig. S1.

impact types and causes should be taken into account when developing drought impact models.

The forecasting skill assessment was carried out using the BSS and ROC approaches to drought hazard and drought impact forecasts. Our results show that, in general, the BSS skill score produces a lower skill than ROC. The ability of the BSS and ROC to distinguish between good and failed forecasts is limited by the number of years in the re-forecast period (Kumar, 2009) and ensemble members (Scaife et al., 2014; Dunstone et al., 2016). The BSS assesses all joint distributions of forecasts and observations based on the ensemble members in probability binary time series. Therefore, BSS is subject to large sampling variability and more sensitive to the details of the datasets (Hamill and Juras, 2006). Conversely, the ROC skill score uses the probability thresholds of the ensemble members. Thus, it may subject to be noisier, but it is insensi-

tive to the particular probabilities themselves and insensitive to bias (Bett et al., 2018). The BSS is also sensitive to rare events, such as extreme wet and dry conditions. Therefore, large sample sizes are required to stabilize the BSS when one evaluates rare events. The reliability of the BSS values of the re-forecast drought impacts is uncertain as these events are subjected to a more skewed distribution, which is not the case for ROC. In the ROC analysis, the success rate in predicting no impact is not part of the goodness-of-fit measure. Therefore, the ROC is not sensitive to biases and does not include reliability. The downside, however, is that it ignores the predicted probability values and the goodness of fit of the model.

## 5 Conclusions

This study evaluates the skills of both drought hazard and impact forecasts, which were developed using observed and re-forecast hydro-meteorological data, and drought impact reports (EDII). Empirical drought impact functions were developed by utilizing logistic regression (LR) and random forest (RF) machine learning approaches to link drought hazards to drought impacts.

Our results show that re-forecast drought hazard indices with longer accumulation periods have higher skill than those with short periods. Longer accumulation periods consist of a longer series of observation data relative to the lead time, which is a caveat for forecasting skill assessment. The higher proportion of observed data will artificially inflate forecast scores. Likely, forecast skill of drought indices with longer accumulation periods also is positively affected by the lower fluctuation/noise in the index time series compared to shorter accumulation periods. Furthermore, higher prediction skill is obtained for hydrological drought hazard indices than for meteorological indices for the same accumulation period, which is associated with the memory represented in initial hydrological conditions and storage in the hydrological system (Sutanto et al., 2020).

The descriptive power of LR and RF to forecast drought impacts was assessed for multiple cases. Both data-driven models show similar skills for predicting drought impact occurrence when compared with observed impact data from 2016 to 2017. However, when the drought impact functions that were trained using all observed impact data and drought hazard indices (1990–2017) were used to predict drought impacts based upon re-forecast data (2002–2010), RF shows higher discriminative ability than LR. Our findings also reveal that models with higher spatial resolution and less lumped impacted sectors (cases 4 and 5) have higher predictive skill than the ones for more generic cases (cases 1–3). For the former cases, drought impacts can be predicted up to 3–4 months ahead with considerable skill.

The forecasting skill of drought impacts greatly depends on the availability of impact data. Data-driven models require sufficient reported impact data to train the model. This study demonstrated that the drought impact functions could not be developed for certain regions and impacted sectors, owing to the lack of reported impacts. Although higher skill was gained for case 5 than case 4, fewer drought impact functions could be developed for case 5. This is a reason that we are in favor of lumping impact data into several impact groups (case 4). In this study, we used all drought hazard indices with all accumulation periods (SPI-x, SPEI-x, and SRI-x) to develop drought impact functions using RF. A further study to develop drought impact functions based on pre-defined drought indices with certain accumulation periods, which are linked to specific impacts, is suggested. For example, drought indices with long accumulation periods are used

to forecast hydrological drought impacts and vice versa for meteorological drought impacts.

*Data availability.* The hydro-meteorological data generated and/or analyzed during this study are freely available from the 4TU Center for Research Data (<https://doi.org/10.4121/uuid:4c6f7f0f-e402-4b6c-a2a2-711bcd224e1f>, Sutanto and van der Weert, 2020).

*Supplement.* The supplement related to this article is available online at: <https://doi.org/10.5194/nhess-20-1595-2020-supplement>.

*Author contributions.* All authors conceived and implemented the research. Data analyses and model output analyses were performed, and all figures were created, by MVDW and SJS. SJS and HAJVL wrote the initial version of the paper. VB contributed to interpreting the results, discussion, and improving the paper.

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

*Special issue statement.* This article is part of the special issue “Recent advances in drought and water scarcity monitoring, modelling, and forecasting (EGU2019, session HS4.1.1/NH1.31)”. It is not associated with a conference.

*Acknowledgements.* The research was supported by the ANYWHERE project (grant agreement no.: 700099), which is funded within the EU’s Horizon 2020 research and innovation program (<http://anywhere-h2020.eu/>, last access: May 2020). The hydro-meteorological output came from the EFAS computational center, which is part of the Copernicus Emergency Management Service (EMS) and Early Warning Systems (EWS). Finally, we acknowledge the Ministry of Science, Research and the Arts of the state of Baden-Württemberg for financing the DRIeR project, which maintains EDII. This research is part of the Wageningen Institute for Environment and Climate Research (WIMEK-SENSE), and it supports the work of the UNESCO EURO FRIEND-Water program and the IAHS Panta Rhei project Drought in the Anthropocene.

*Financial support.* This research has been supported by the H2020 (grant no. 700099).

*Review statement.* This paper was edited by Athanasios Loukas and reviewed by two anonymous referees.

## References

- Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., and Pappenberger, F.: Skilful seasonal forecasts of streamflow over Europe?, *Hydrol. Earth Syst. Sci.*, 22, 2057–2072, <https://doi.org/10.5194/hess-22-2057-2018>, 2018.
- Bachmair, S., Svensson, C., Hannaford, J., Barker, L. J., and Stahl, K.: A quantitative analysis to objectively appraise drought indicators and model drought impacts, *Hydrol. Earth Syst. Sci.*, 20, 2589–2609, <https://doi.org/10.5194/hess-20-2589-2016>, 2016.
- Bachmair, S., Svensson, C., Prosdocimi, I., Hannaford, J., and Stahl, K.: Developing drought impact functions for drought risk management, *Nat. Hazards Earth Syst. Sci.*, 17, 1947–1960, <https://doi.org/10.5194/nhess-17-1947-2017>, 2017.
- Bartholmes, J., Thielen, J., and Kalas, M.: Forecasting medium-range flood hazard on European scale, *Georisk*, 2, 181–186, <https://doi.org/10.1080/17499510802369132>, 2008.
- Bett, P., Thornton, H., and Troccoli, A.: Skill assessment of energy-relevant climate variables in a selection of seasonal forecast models, Report using final data sets, ECEM Deliverable D2.2.1 v2, ECMWF Copernicus Report, ECMWF, Reading, UK, 2018.
- Blauhut, V., Gudmundsson, L., and Stahl, K.: Towards pan-European drought risk maps: quantifying the link between drought indices and reported drought impacts, *Environ. Res. Lett.*, 10, 014008, <https://doi.org/10.1088/1748-9326/10/1/014008>, 2015.
- 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, *Hydrol. Earth Syst. Sci.*, 20, 2779–2800, <https://doi.org/10.5194/hess-20-2779-2016>, 2016.
- Breiman, L.: Random forests, *Mach. Learn.*, 45, 5–32, 2001.
- Brier, G. W.: Verification of forecasts expressed in terms of probability, *Mon. Weather Rev.*, 78, 1–3, 1950.
- Burek, P., van der Knijff, J., and Roo, A. D.: LISFLOOD distributed water balance and flood simulation model, JRC Tech. Report, JRC – Joint Research Centre, Ispra, Italy, <https://doi.org/10.2788/24719>, 2013.
- Camia, A., Durrant, T., and San-Miguel-Ayanz, J.: The European fire database technical specifications and data submission, Executive Report, Joint Research Center, Ispra, Italy, <https://doi.org/10.2788/2175>, 2014.
- Carrão, H., Naumann, G., and Barbosa, P.: Mapping global pattern of drought risk: An empirical framework based on sub-national estimates of hazard, exposure and vulnerability, *Global Environ. Change*, 39, 108–124, <https://doi.org/10.1016/j.gloenvcha.2016.04.012>, 2016.
- Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N., Andrews, M., and Knight, J.: Skilful predictions of the winter North Atlantic Oscillation one year ahead, *Nat. Geosci.*, 9, 809–814, <https://doi.org/10.1038/ngeo2824>, 2016.
- EEA: Climate change and water adaptation issues, EEA Technical Report No. 2/2007, European Environmental Agency, Copenhagen, Denmark, 2007.
- EEA: Urban adaptation to climate change in Europe Challenges and opportunities for cities together with supportive national and European policies, Report No. 2/2012, European Environment Agency, Copenhagen, Denmark, 2012.
- Elshorbagy, A., Corzo, G., Srinivasulu, S., and Solomatine, D. P.: Experimental investigation of the predictive capabilities of data driven modeling techniques in hydrology – Part I: Concepts and methodology, *Hydrol. Earth Syst. Sci.*, 14, 1931–1941, <https://doi.org/10.5194/hess-14-1931-2010>, 2010.
- Eurostat: Regions in the European union: Nomenclature of territorial units for statistics nuts 2010/eu-27, Eur. Union, Publications Office of the European Union, Luxembourg, <https://doi.org/10.2785/15544>, 2011.
- Evans, J. S., Murphy, M. A., Holden, Z. A., and Cushman, S. A.: Modeling Species Distribution and Change Using Random Forest, in: Predictive Species and Habitat Modeling in Landscape Ecology, edited by: Drew, C. A., Wiersma, Y. F., and Huettmann, F., Springer, New York, NY, 313 pp., 2011.
- Feyen, L. and Dankers, R.: Impact of global warming on streamflow drought in Europe, *J. Geophys. Res.*, 114, D17116, <https://doi.org/10.1029/2008JD011438>, 2009.
- Forzieri, G., Feyen, L., Rojas, R., Flörke, M., Wimmer, F., and Bianchi, A.: Ensemble projections of future streamflow droughts in Europe, *Hydrol. Earth Syst. Sci.*, 18, 85–108, <https://doi.org/10.5194/hess-18-85-2014>, 2014.
- Fung, K. F., Huang, Y. F., Koo, C. H., and Soh, Y. W.: Drought forecasting: a review of modeling approaches 2007–2017, *J. Water Clim. Change*, <https://doi.org/10.2166/wcc.2019.236>, in press, 2019.
- Georgi, B., Swart, R., Marinova, N., Van Hove, B., Jacobs, C., and Klostermann, J.: Urban adaptation to climate change in Europe: Challenges and opportunities for cities together with supportive national and European policies, Tech. rep., EEA, available at: <https://www.eea.europa.eu/publications/urban-adaptation-to-climate-change> (last access: May 2020), 2012.
- Gudmundsson, L., Rego, F., Rocha, M., and Seneviratne, S. I.: Predicting above normal wildfire activity in southern Europe as a function of meteorological drought, *Environ. Res. Lett.*, 9, 084008, <https://doi.org/10.1088/1748-9326/9/8/084008>, 2014.
- Hamill, T. M. and Juras, J.: Measuring forecast skill: is it real skill or is it the varying climatology?, *Q. J. Roy. Meteorol. Soc.*, 132, 2905–2923, <https://doi.org/10.1256/qj.06.25>, 2006.
- Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning, Springer, New York, NY, 2009.
- Hayes, M., Svoboda, M., Wall, N., and Widhalm, M.: The Lincoln declaration on drought indices, universal meteorological drought index recommended, *B. Am. Meteorol. Soc.*, 485–488, <https://doi.org/10.1175/2010BAMS3103.1>, 2011.
- Kirasich, K., Smith, T., and Sadler, B.: Random forest vs logistic regression: binary classification for heterogeneous datasets, *SMU Data Sci. Rev.*, 1, 9, 2018.
- Kreibich, H., Blauhut, V., Aerts, J. C. J. H., Bouwer, L. M., Van Lanen, H. A. J., Mejia, A., Mens, M., and Van Loon, A. F.: How to improve attribution of changes in drought and flood impacts, *Hydrolog. Sci. J.*, 64, 1–18, <https://doi.org/10.1080/02626667.2018.1558367>, 2019.
- Kuhn, M.: Classification and regression training, package ‘caret’, CRAN R Project, available at: <https://cran.r-project.org/web/packages/caret/caret.pdf> (last access: April 2020), 2019.
- Kumar, A.: Finite samples and uncertainty estimates for skill measures for seasonal prediction, *Mon. Weather Rev.*, 137, 2622–2631, <https://doi.org/10.1175/2009mwr2814.1>, 2009.

- Lackstrom, K., Brennan, A., Ferguson, D., Crimmins, M., Darby, L., Dow, K., Ingram, K., Meadow, A., Reges, H., Shafer, M., and Smith, K.: The Missing Piece: Drought Impacts Monitoring, in: Workshop report produced by the Carolinas Integrated Sciences & Assessments program and the Climate Assessment for the Southwest, 5–6 March 2013, Tucson, AZ, 1–23, available at: <https://www.climas.arizona.edu/publication/report/missing-piece-drought-impacts-monitoring> (last access: March 2020), 2013.
- Liaw, A. and Wiener, M.: Classification and regression by random forest, *R News*, 2, 18–22, 2002.
- Mason, I.: A model for assessment of weather forecasts, *Aust. Meteorol. Mag.*, 30, 291–303, 1982.
- Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, *Weather Forecast.*, 14, 713–725, [https://doi.org/10.1175/1520-0434\(1999\)014<0713:CPROCA>2.0.CO;2](https://doi.org/10.1175/1520-0434(1999)014<0713:CPROCA>2.0.CO;2), 1999.
- McKee, T. B., Doesken, N. J., and Kleist, J.: The relationship of drought frequency and duration to time scales, in: Vol. 17, Proceedings of the 8th Conference on Applied Climatology, American Meteorological Society Boston, MA, 179–183, 1993.
- Mishra, A. K. and Singh, V. P.: A review of drought concepts, *J. Hydrol.*, 391, 202–216, <https://doi.org/10.1016/j.jhydrol.2010.07.012>, 2010.
- Mount, N. J., Maier, H. R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., and Abrahart, R. J.: Data-driven modelling approaches for socio-hydrology: opportunities and challenges within the Panta Rhei Science Plan, *Hydrolog. Sci. J.*, 61, 1192–1208, <https://doi.org/10.1080/02626667.2016.1159683>, 2016.
- 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, *Hydrol. Earth Syst. Sci.*, 18, 1591–1604, <https://doi.org/10.5194/hess-18-1591-2014>, 2014.
- Naumann, G., Spinoni J., Vogt, J., and Barbosa, P.: Assessment of drought damages and their uncertainties in Europe, *Environ. Res. Lett.*, 10, 124013, <https://doi.org/10.1088/1748-9326/10/12/124013>, 2015.
- Pappenberger, F., Thielen, J., and Del Medico, M.: The impact of weather forecast improvements on large scale hydrology: analyzing a decade of forecasts of the European Flood Alert System, *Hydrol. Process.*, 25, 1091–1113, <https://doi.org/10.1002/hyp.7772>, 2011.
- Park, S., Im, J., Jang, E., and Rhee, J.: Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions, *Agr. Forest Meteorol.*, 216, 157–169, <https://doi.org/10.1016/j.agrformet.2015.10.011>, 2015.
- Peng, C.-Y. J., Lee, K. L., and Ingersoll, G. M.: An introduction to logistic regression analysis and reporting, *J. Educ. Res.*, 96, 3–14, <https://doi.org/10.1080/00220670209598786>, 2002.
- Poljanšek, K., Marin Ferrer, M., De Groeve, T., and Clark, I.: Science for disaster risk management 2017: knowing better and losing less, executive summary, JRC 102482, EUR 28034 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-79-60679-3, <https://doi.org/10.2788/842809>, 2017.
- Pozzi, W., Sheffield, J., Stefanski, R., Cripe, D., Pulwarty, R., Vogt, J. V., Heim Jr., R. R., Brewer, M. J., Svoboda, M., Westerhoff, R., van Dijk, A. I. J. M., Lloyd-Hughes, B., Pappenberger, F., Werner, M., Dutra, E., Wetterhall, F., Wagner, W., Schubert, S., Mo, K., Nicholson, M., Bettio, L., Nunez, L., van Beek, R., Bierkens, M., de Goncalves, L. G. G., de Mattos, J. G. Z., and Lawford, R.: Toward global drought early warning capability: Expanding international cooperation for the development of a framework for monitoring and forecasting, *B. Am. Meteorol. Soc.*, 94, 776–785, <https://doi.org/10.1175/BAMS-D-11-00176.1>, 2013.
- Rhee, J. and Im, J.: Meteorological drought forecasting for ungauged areas based on machine learning: Using long-range climate forecast and remote sensing data, *Agr. Forest Meteorol.*, 237–238, 105–122, <https://doi.org/10.1016/j.agrformet.2017.02.011>, 2017.
- Ripley, B., Venables, B., Bates, D. M., Hornik, K., Gebhardt, A., and Firth, D.: Package “mass”, CRAN Repos., available at: <http://www.stats.ox.ac.uk/pub/MASS4/> (last access: April 2020), 2019.
- Scaife, A. A., Arribas, A., Blockley, E., Brookshaw, A., Clark, R. T., Dunstone, N., Eade, R., Fereday, D., Folland, C. K., Gordon, M., Hermanson, L., Knight, J. R., Lea, D. J., MacLachlan, C., Maidens, A., Martin, M., Peterson, A. K., Smith, D., Vellinga, M., Wallace, E., Waters, J., and Williams, A.: Skillful long-range prediction of European and North American winters, *Geophys. Res. Lett.*, 41, 2514–2519, <https://doi.org/10.1002/2014gl059637>, 2014.
- Sepulcre-Canto, G., Horion, S., Singleton, A., Carrao, H., and Vogt, J.: Development of a combined drought indicator to detect agricultural drought in Europe, *Nat. Hazards Earth Syst. Sci.*, 12, 3519–3531, <https://doi.org/10.5194/nhess-12-3519-2012>, 2012.
- Sheffield, J., Wood, E. F., Chaney, N., Guan, K., Sadri, S., Yuan, X., Olang, L., Amani, A., Ali, A., Demuth, S., and Ogallo, L.: A drought monitoring and forecasting system for sub-Saharan African water resources and food security, *B. Am. Meteorol. Soc.*, 95, 861–882, <https://doi.org/10.1175/BAMS-D-12-00124.1>, 2014.
- Shukla, S. and Wood, A. W.: Use of a standardized runoff index for characterizing hydrologic drought, *Geophys. Res. Lett.*, 35, L02405, <https://doi.org/10.1029/2007GL032487>, 2008.
- Smith, P., Pappenberger, F., Wetterhall, F., Thielen, J., Krzeminski, B., Salamon, P., Muraro, D., Kalas, M., and Baugh, C.: On the operational implementation of the European Flood Awareness System (EFAS), ECMWF Tech. Memorandum 778, ECMWF, Reading, UK, 1–34, 2016.
- Solomatine, D. P. and Shrestha, L.: A novel method to estimate model uncertainty using machine learning techniques, *Water Resour. Res.*, 45, W00B11, <https://doi.org/10.1029/2008WR006839>, 2009.
- Spinoni, J., Naumann, G., Vogt, J. V., and Babosa, P.: The biggest drought events in Europe from 1950 to 2012, *J. Hydrol.: Reg. Stud.*, 3, 509–524, <https://doi.org/10.1016/j.ejrh.2015.01.001>, 2015.
- Stagge, J. H., Kohn, I., Tallaksen, L. M., and Stahl, K.: Modeling drought impact occurrence based on meteorological drought indices in Europe, *J. Hydrol.*, 530, 37–50, <https://doi.org/10.1016/j.jhydrol.2015.09.039>, 2015.
- Stahl, K., Kohn, I., Blauhut, V., Urquijo, J., De Stefano, L., Acácio, V., Dias, S., Stagge, J. H., Tallaksen, L. M., Kampragou, E., Van Loon, A. F., Barker, L. J., Melsen, L. A., Bifulco, C., Musolino, D., de Carli, A., Massarutto, A., Assi-

- macopoulos, D., and Van Lanen, H. A. J.: Impacts of European drought events: insights from an international database of text-based reports, *Nat. Hazards Earth Syst. Sci.*, 16, 801–819, <https://doi.org/10.5194/nhess-16-801-2016>, 2016.
- Sutanto, S. J. and van der Weert, M.: Supporting data for ‘Skill of large-scale seasonal drought impact forecasts’, <https://doi.org/10.4121/uuid:4c6f7f0f-e402-4b6c-a2a2-711bcd224e1f>, 2020.
- Sutanto, S. J., Van Lanen, H. A. J., Wetterhall, F., and Llort, X.: Potential of pan-European seasonal hydro-meteorological drought forecasts obtained from a Multi-Hazard Early Warning System, *B. Am. Meteorol. Soc.*, 101, 368–393, <https://doi.org/10.1175/BAMS-D-18-0196.1>, 2019a.
- Sutanto, S. J., Van der Weert, M., Wanders, N., Blauhut, V., and Van Lanen, H. A. J.: Moving from drought hazard to impact forecasts, *Nat. Commun.*, 10, 4945, <https://doi.org/10.1038/s41467-019-12840-z>, 2019b.
- Sutanto, S. J., Wetterhall, F., and Van Lanen, H. A. J.: Hydrological drought forecasts outperform meteorological drought forecasts, *Environ. Res. Lett.*, <https://doi.org/10.1088/1748-9326/ab8b13>, in press, 2020.
- Tallaksen, L. M. and Van Lanen, H. A. J.: Hydrological drought: processes and estimation methods for streamflow and groundwater, in: *Developments in Water Science*, vol. 48, Elsevier Science B.V., Amsterdam, the Netherlands, 2004.
- Thielen, J., Bartholmes, J., Ramos, M.-H., and Roo, A. D.: The European Flood Alert System – Part I: concept and development, *Hydrol. Earth Syst. Sci.*, 13, 125–140, <https://doi.org/10.5194/hess-13-125-2009>, 2009.
- Trambauer, P., Werner, M., Winsemius, H. C., Maskey, S., Dutra, E., and Uhlenbrook, S.: Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa, *Hydrol. Earth Syst. Sci.*, 19, 1695–1711, <https://doi.org/10.5194/hess-19-1695-2015>, 2015.
- Trigila, A., Ladanza, C., Esposito, C., and Scarascia-Mugnozza, G.: Comparison of Logistic Regression and Random Forest techniques for shallow landslide susceptibility assessment in Giampileri (NE Sicily, Italy), *Geomorphology*, 249, 119–136, <https://doi.org/10.1016/j.geomorph.2015.06.001>, 2015.
- UNDP: Summary Human Development Report, Beyond Scarcity: Power, Poverty and the Global Water Cycle, Palgrave MacMillan, New York, 2006.
- Van Der Knijff, J.: LISVAP: Evaporation Pre-Processor for the LISFLOOD Water Balance and Flood Simulation Model, EUR 22639 EN/2, JRC – Joint Research Centre, Ispra, Italy, 2008.
- Van Der Knijff, J. M., Younis, J., and De Roo, A. P. J.: Lisflood: a gis-based distributed model for river basin scale water balance and flood simulation, *Int. J. Geogr. Inform. Sci.*, 24, 189–212, <https://doi.org/10.1080/13658810802549154>, 2010.
- Van Hateren, T., Sutanto, S. J., and Van Lanen, H. A. J.: Evaluating uncertainty and robustness of seasonal meteorological and hydrological drought forecasts at the catchment scale-case Catalonia (Spain), *Environ. Int.*, 133, 105206, <https://doi.org/10.1016/j.envint.2019.105206>, 2019.
- Van Loon, A. F.: Hydrological drought explained, *WIREs Water*, 2, 359–392, <https://doi.org/10.1002/wat2.1085>, 2015.
- Vicente-Serrano, S. M., Begueria, S., and López-Moreno, J. I.: A multi-scalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index-SPEI, *J. Climate*, 23, 1696–1718, <https://doi.org/10.1175/2009JCLI2909.1>, 2010.
- Vitart, F.: Evolution of ECMWF sub-seasonal forecast skill scores, *Q. J. Roy. Meteorol. Soc.*, 140, 1889–1899, <https://doi.org/10.1002/qj.2256>, 2014.
- Wilhite, D. A.: ed. Drought: A Global Assessment, in: *Routledge Hazards and Disasters Series*, vol. I & II, Routledge, London, 2000.
- Wilhite, D. A. and Vanyarkho, O. V.: Drought: Pervasive Impacts of a Creeping Phenomenon, in: *Drought: A Global Assessment*, Vol. I, chap. 18, edited by: Wilhite, D. A., London, Routledge, 245–255, 2000.
- Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., and Smith, G. M.: *GLM and GAM for Absence–Presence and Proportional Data*, Springer, New York, 245–259, 2009.