Application of machine learning to model wetland inundation patterns across a large semi-arid floodplain

S. Shaeri Karimi¹*, N. Saintilan¹, L. Wen² and R. Valavi³

¹Department of Earth and Environmental Sciences, Faculty of Science and Engineering, Macquarie University, Sydney, NSW 2109, Australia.
²Department of Planning, Industry and Environment, NSW Australia.
³School of Biosciences, University of Melbourne, Parkville, Vic. 3010, Australia.

*Corresponding author: Sara Shaeri Karimi (sara.shaeri-karimi@students.mq.edu.au)

Key Points:

- A machine learning technique (Random Forest) was successfully applied to predict daily floodplain inundation at a fine resolution of 30 m.
- The down-sampling method effectively improved model performance
- Topography and 30-day cumulative discharge are the most influential parameters predicting inundation occurrence.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1029/2019WR024884
Abstract

Inundation is a primary driver of floodplain ecology. Understanding temporal and spatial variability of inundation patterns is critical for optimum resource management, particularly in striking an appropriate balance between environmental water application and extractive use. Nevertheless, quantifying inundation at the fine resolution required of ecological modeling is an immense challenge in these environments. In this study, random forest, a machine learning technique, was implemented to predict the inundation pattern in a section of the Darling River Floodplain, Australia at a spatial scale of 30 m and daily temporal resolution. The model achieved very good performance with an average accuracy of 0.915 based on the Area under the Receiver Operating Characteristic (AUCROC) curve over 10 runs of the model in testing data sets. Six variables explained 70% of the total contribution to inundation occurrence, with the most influential being landscape shape (local deviation from global mean elevation), elevation-weighted distance to the river, the magnitude of river flow (10-day and 30-day accumulated river discharge), local rainfall and soil moisture. This approach is applicable to other floodplains across the world where understanding of fine-scale inundation pattern is for operational ecological management and scenario testing.

1 Introduction

Water extraction for agricultural and urban water use has compromised the river flow available to maintain ecological processes and ecological assets, both in-stream (Randklev et al., 2018) and on floodplains (Chiu et al., 2017). The decline of riverine ecosystems in semi-arid landscapes has been well documented (Hauer & Lorang 2004; Glenn et al., 2017), and prompted the call for revision of the proportional allocation of water for extractive use and the maintenance of environmental goods and services. In this context, the application of predictive hydrological models linking flow to environmental outcomes is a crucial component of optimal water planning.

In Australia, the world’s driest inhabited continent, this tension has been played out at a large scale in the nation’s most agriculturally important river system, the Murray Darling Basin (MDB). The MDB covers approximately 1 million square kilometres, responsible for more than 30 percent of Australia’s agricultural production (Saintilan & Overton 2010). A long history of water diversion and allocation under license for agricultural production led to a situation in which levels of abstraction were unsustainable, brought to a head during a prolonged drought (2000-2009) in which the ecological condition of rivers and wetlands listed as internationally significant under the Ramsar convention declined to critical levels (Rogers & Ralph 2010). In response, the Commonwealth Government of Australia promulgated the Water Act (2009), which required the development of a plan for a more sustainable balance between extractive use and environmental outcomes (the Murray Darling Basin Plan), and the provision of more than $3
billion to purchase back water licenses for environmental use, and for water efficiency measures across the basin (MDBA, 2018).

This considerable level of investment in environmental water, of which the Murray Darling Basin Plan and associated water buyback is one example among many (Landry, 1998; Qureshi et al. 2010), that brings into focus the importance of managing water with optimal efficiency, both in irrigated agriculture and in environmental application (Arthington et al., 2006; Saintilan et al., 2013). The strategic use of environmental water requires the definition of flow targets under a range of climatic conditions, including targets for the inundation of ecological assets situated on floodplains. These targeted assets include floodplain forests, wetlands and ephemeral water bodies (e.g. cut-off meanders or billabongs), that provide ecological services including habitat provision (e.g. waterbird breeding, amphibian, turtle and ephemeral fish habitat) and the flux of carbon from the floodplain to the river (Kobayashi et al., 2011). The accurate assessment and modeling of floodplain inundation is thereby becoming an important operational tool in river management, and a key reporting requirement in documenting outcomes of investment in environmental water.

The measurement and characterization of floodplain inundation are technically challenging (Teng et al., 2017). Whereas instream flows are relatively easily characterized by flow gauges, or water stage levels and developed stage discharge curves, the progression of water in overbank flow can only realistically be measured by remote sensing approaches. Though a range of remote sensing platforms have been used to identify water on floodplains (Hess et al., 2003; Prigent et al., 2001), including active sensors (Mitchell et al., 2015), the most common tool has been Landsat TM (Frazier et al., 2003; Thomas et al., 2011), which provides 30 m resolution, for a 16-day return period, and the benefit of a long archival record (extending back to 1986). The Landsat TM archive has been applied to comprehend the surface water extent dynamics at scales ranging from floodplain (Thomas et al., 2015) to subcontinental (Tulbere et al. 2016; Heimhuber et al., 2017; Senay et al., 2017) to continental (Mueller et al., 2016) and global (Donchyts et al., 2016; Pekel et al., 2016).

In spite of the availability of the Landsat TM archive and derived products, considerable difficulties remain in their application for modeling floodplain inundation. First, the 16-day return period presents only a partial record of inundation for flood events, which may occur over a period of days to weeks. The temporal resolution of Landsat TM is further compromised by cloud-affected imagery, which might often be co-incident with the timing of overbank flows. Further, inundation may occur underneath floodplain vegetation, making detection difficult. At global scale, Pekel et al. (2016) detected seasonality, occurrences and changes in surface water based on more than 3 million Landsat satellite images over a 32-year period at 30-meter resolution on a monthly time scale, and the dataset is freely available. Although the data are satisfactorily accurate for large water bodies, anecdotal evidence suggests the usefulness for arid and semi-arid floodplains may be limited. For example, we found large discrepancy at vegetated swamps and anabranching channels in Murrumbidgee Valley National Park when compared with
field observations. In addition, the accuracy in high-latitude regions may be compromised due to snow, ice, persistent cloud cover, sediment suspension or poor data quality (Nitze et al., 2017). The snap shots of inundation derived from the Landsat TM archive are in many contexts insufficient to meet three important goals of floodplain inundation mapping: (i) to properly characterize inundation at the event scale (Thomas et al., 2015); (ii) to provide insight into the inundation patterns associated with different ecological habitats and processes on the floodplain (Broich et al., 2018), and (iii) to provide the time-series data of sufficient temporal resolution required to understand the implication of historic diversions or changing climatic conditions on floodplain inundation (Ren et al., 2011). Using the downscaling algorithm, Aires et al. (2017) developed a global monthly time series of fine resolution (~90 m) inundation from the course ~25 km GIEMS (the Global Inundation Extent from Multi-Satellites). Despite the high accuracy in wet areas, large discrepancies were found for semi-arid areas (Aires et al. 2017).

The derivation of a longer time series of inundation at a higher temporal resolution than that permitted by Landsat TM observations requires the application of either physics-based hydrodynamic models or data-based statistical models (Teng et al., 2017). As for the former, a combination of hydrodynamic and hydrological modeling approaches has been applied to predict the progression of water across the floodplain under a range of hydrological conditions. For example, Wen et al. (2013) developed 2D floodplain hydrodynamic models for the Macquarie Marshes based on LiDAR-derived elevation models, soil hydrometric and floodplain vegetation characteristics, calibrated against Landsat inundation measures. However, these Mike-Flood models are computationally intensive and impractical for the modeling of a longer time-series of inundation. They are unviable for areas larger than 1,000 km² for fine simulation (Teng et al., 2017), and require input data on roughness that may be unavailable or subject to alteration within model runs. For this purpose, relationships between stage height and inundation were extracted from the hydrodynamic model and applied in a wetland hydrological model, which represented inundation extent and duration at the scale of wetland eco-geomorphological compartments. These models were further developed and applied to characterize floodplain vegetation distribution by Sandi et al. (2015, 2018).

The characterization of floodplain inundation using hydrodynamic models is complicated by a number of factors unique to the floodplain. The primary boundary condition of elevation is made problematic by the comparative lack of relief and errors in relief estimation from LiDAR (including, for example, the absence of ground returns in thick vegetation). System complexity is increased by the high spatial variability in resistance resulting from the complexity of floodplain vegetation, and the propensity of this vegetation to alter at the event scale. Losses to groundwater vary considerably in relation to antecedent hydrological conditions. Finally, the model domain is difficult to constrain in the context of inputs and losses, not only from the river channel, but from elsewhere on the floodplain. Given the challenges in developing hydrodynamic models with appropriate temporal and spatial resolution and scale, several attempts have been made to predict floodplain inundation using data-based statistical approaches (see Heimhuber et al., 2015).
Statistical inundation models used to be of coarse spatial and temporal resolution due to data availability (Smith, 1997), but recent advances in remote sensing greatly improve the situation.

The aim of this study was to develop a spatially explicit predictive model of floodplain inundation using machine-learning techniques. Machine learning uses algorithms to build analytical models, helping the computer “learn” directly and dynamically from data (Witten et al. 2016), in contrast to models that follow pre-programmed rules. The approach can be applied to rapidly identify patterns and tendencies in big datasets, and facilitate analysis for diverse problems in many disciplines. This study used Random Forest (RF), a tree-based machine learning technique to determine the inundation patterns in a large semi-arid floodplain using a combination of topographic, hydrological and climatic variables. RF has been successfully applied in a wide range of earth science and ecology studies such as, habitat modeling (Cutler, 2007; Peters, 2007), species distribution and identification (Evan & Cushman, 2009; Armitage, 2010; D. Rosa et al., 2016; Valavi et al., 2019a), remote sensing (Belgiu & Dragut, 2016), risk assessment (Wang et al., 2015; Lee et al., 2017; Feng et al., 2015) and water resource studies (Booker and Snelder, 2012; Alvarez-Cabria et al., 2016; Tulbure et al., 2016; González-Ferreras & Barquin, 2017). The target location, a section of the Darling River Floodplain in New South Wales, is of interest because of the rapidly expanding irrigation demand in the catchment, contentious cross-border water management issues, and the lack of detailed hydrologic and ecological investigations compared to floodplains further south despite the existence of high quality elevation models, extensive river flow and climate records, cloud-free Landsat flood imagery and detailed vegetation mapping.

2 Materials and Methods

2.1 Study area

The study area is located in the north-west of New South Wales, Australia, and spans a 118 km reach of the Darling River, initiating approximately 50 km downstream of river gauging station at Bourke (Station Number: 425003). The area covers the Warrego river junction through to Louth weir 21 (Figure 1). The studied reach covers an area of 1,200 km², with terrain elevation ranging between 80 and 145 m above Australian Height Datum. The river channel is between 60 to 80 meters wide and up to 25 meters deep (Crown Lands and Water Division, 2017). The study area is a reach of the largely unregulated Barwon-Darling system in NSW with no major water storage structures to regulate discharge in the system (MDBA, 2011). The long-term average river flow is estimated to be 8,446 ML/day at Bourke. When discharge exceeds 14,000 ML/day, up to 77% of this area are inundated (Brennan et al. 2002). Flooding in this area is primarily confined to the right bank (in the direction of flow) of the Darling River with floodwaters crossing the Warrego River in the area upstream of its confluence with the Darling during flood events in the latter river (NSW DPI Water, 2017).

The area is considered to have extreme climatic variability with low rainfall (Thoms & Sheldon, 2000) throughout the year and a meagre annual average of just 305.6 mm for the period
of 1998-2018 (BOM, 2018). The average daily minimum temperature is 4.1 °C in July and the average daily maximum temperature of 37.3 °C in January (BOM, 2018). The area includes 320 km² of Toorale National Park and State Conservation Area; however, grazing is the main land use across the study reach (Crown Lands and Water Division, 2017).

2.2 Modeling approach

The study was performed in four steps. In the first step, a dataset of dependent variables (response) and independent variables (predictors), describing several environmental characteristics of the study area was created from various sources in the format of raster files. In the second step, 10,000 points from the study area were randomly sampled from the stacked raster, and subsequently transformed to a data frame to serve as input for the model. Fifteen independent variables were selected amongst the initial set of 36 predictors to avoid multicollinearity.

In step three, the dataset was balanced. The entire dataset is highly biased to non-inundated (class 0, amounted for 93% of the data points) reflecting the regional semi-arid climate. The unbalanced nature of the dataset may lead to low predictive power. Thus, we used the down-sampling method (Chen et al., 2004) to balance the classes. The RF model was subsequently trained using the balanced dataset. Finally, the trained RF model was applied to the test dataset and model performance evaluated based on a number of different performance metrics. The observed inundation was subsequently compared with the model prediction to assess the models’ ability to forecast the inundation extent. An expansion of the modeling approach is discussed in the following sections.

2.2.1 Dependent variable

A set of 74 binary (inundated/non-inundated) Landsat derived rasters describing the inundation extent of the study area with a spatial resolution of 30m × 30m was provided by the NSW Office of Environment and Heritage covering years from 1988 to 2016 which are archived on the Open Science Framework (Available at https://doi.org/10.17605/OSF.IO/PCNJQ). The temporal resolution of the inundation rasters varied due to the selection of dates to obtain a high percentage of inundation and low percentage of cloud cover (Powell et al., 2017).

2.2.2 Independent variables

This article is protected by copyright. All rights reserved.
Independent variables representing several topographic, climatic, and hydrologic attributes were extracted from existing databases provided by several regional organizations. An initial set of 36 predictors with potential influence on flood inundation were selected as follows:

- Topographic variables (Table 1) were derived from the 5 m digital elevation model (DEM), (Available at https://data.gov.au/dataset/ed99d012-3e31-436c-b42b-97a4415a1d96) using “Geomorphometry and Gradient Metrics” toolbox (Evan et al., 2014) in ArcGIS 10.5.

- The daily rainfall data from 1988 to 2016 were provided by the Bureau of Meteorology (BOM) as gridded 5×5 km data. Additionally, the BOM provided runoff and soil moisture data to the same resolution, from the Australian Water Resources Assessment Landscape (AWRA-L v5.0) modeling system, which is a grid based distributed water balance model that is conceptualized as a small catchment (Viney et al., 2015). The data are available through the Australian Bureau of Meteorology data portal (http://www.bom.gov.au/water/landscape) back to 2005 only although a dataset stretching back to 1911 can be provided upon request.


- The cumulative rainfall (Santos & Fragoso, 2016; Avila et al., 2015; Tramblay et al., 2010), runoff, soil moisture and discharge for 5, 10, 15, 20, 25 and 30 days prior to the captured Landsat inundation scene.

All the mentioned variables were resampled to 30×30 m resolution with bilinear interpolation in ArcGIS for the subsequent application of the RF model. The bilinear resampling computes the new value of a cell based on a weighted distance average of the four nearest input cells. In the next step, all independent and dependent variable rasters were stacked together and 10,000 points were randomly extracted using the function “randomPoints” of the package “dismo” (Hijman et al., 2017), to create a dataset representing the environmental conditions of the whole study area.

Prior to modeling, multi-collinearity was assessed amongst the predictors using Variance Inflation Factors (VIF). According to diverse literature, variables with VIF value greater than 5 indicate a potential problematic multi-collinearity (Rogerson, 2001; Zuur et al., 2009 and James et al., 2013), and as such were removed from the dataset. Finally, fifteen independent variables (Table 1), describing several environmental attributes that were potentially important for modeling the inundation pattern of the studied area were selected.

Table 1
2.2.3 Random Forest Model

The Random Forest (RF) technique is a non-parametric algorithm which generates many classification and regression trees that are aggregated to compute a final prediction. Instead of using all the predictor variables, each tree is fitted using a small, randomly selected subset of features. RF uses a bootstrap sampling, (using about 2/3 of training samples), to grow each tree. Thus, about one-third of the samples are left out-of-bag (OOB), to assess the prediction performance of the Random Forest algorithm (Bierman, 2001; Breiman et al., 1984). RF has been cited as having many advantages, including high classification accuracy, estimation of variable importance, ability to cope with high data dimensionality and multi-collinearity, less sensitive to overfitting and robust to outliers and noise, doesn’t require any assumptions of data distributions (Breiman, 2001; Cutler et al., 2007; Belgiu & Dragut, 2016). RF models were fitted using the “randomForest” package (Liaw & Wiener, 2002) in R environment (R Core Team, 2018).

2.2.3.1 Model Tuning

RF model performance can be optimized for the number of trees (ntree) to grow and the number of predictors used at each node (mtry). In this study, for the parameter mtry, several values were considered varying from 2 to the whole number of predictors according to Kuhn and Johnson (2013). The model with the best performance (minimum OOB Error) with mtry= 8 was finally chosen (Figure 2, left).

As for ntree, the optimal value should make the OOB error of RF converge (Feng et al., 2015). The Number of trees (ntree) was initiated at 500. Based on the plot of ntree vs. OOB error (Figure 2, right), the OOB error was stable when ntree reached 250. To avoid an increase in computation cost (Oshiro et al., 2012) the number of trees was set to 250 as shown in the previous studies mentioned by Lin et al., 2012.

Figure 2

2.2.4 Imbalanced Data

As mentioned earlier, the dataset was highly imbalanced, due to the domination of the non-inundated class (669,746 of a total of 720,612 observations), whilst the inundated class comprised of only 50,866 values. RF uses bootstrap sampling to grow the trees and in the case of an unbalanced data set, it is highly probable that a bootstrap sample contains few or even none of the minority class, leading to poor accuracy in the minority class prediction (Chen et al., 2004). Several techniques have been proposed by different researchers to handle the imbalanced data
set. In this study internal down-sampling, or Balanced Random Forest, was applied to tackle the class imbalance issue. For each tree at each iteration, balanced bootstrap was sampled within a stratified sampling, as such, a bootstrap was chosen from the minority class and then the same number of cases drawn with replacement from the majority class (Chen et al., 2004; Kuhn & Johnson, 2013).

2.2.5 Model Evaluation

Since the model was aimed at predicting into a new time period, block Cross Validation (CV) was used to evaluate model performance (Roberts et al., 2017; Valavi et al., 2019b), which is the same strategy used by Gudmundsson and Seneviratne (2015) and Meyer et al. (2018), respectively named “Cross Validation in Time” and “Leave-Time-Out Cross Validation”. 10-fold time block CV was applied to partition the data set into training and testing subsets. In block CV, for each trial, information of 7 or 8 days was withheld from the dataset for subsequent validation of the learned model. Furthermore, by applying block CV instead of random CV, each fold contains approximately the same percentage of samples of each class as the original dataset in order to have uniform subsamples.

The Area under the Receiver Operating Characteristic (AUC\textsubscript{ROC}) curve has been used as the main threshold-independent performance measure. The ROC curve is a graphical representation of the trade-off between increasing Sensitivity (proportion of all inundated observations correctly predicted) and decreasing Specificity (proportion of all non-inundated observations correctly predicted) for every possible threshold (Chen et al., 2004; Valdes-Donoso et al., 2017). As this model is applied for practical purposes, converting the continuous probability values to categorical values of inundated/non-inundated by selecting appropriate threshold can be useful to evaluate the model performance based on threshold-dependent parameters.

Sensitivity-Specificity sum maximizer approach was applied to identify the best prediction threshold (Liu et al., 2005; Jiménez-Valverde & Lobo, 2007) and the model performance was also evaluated using Model Sensitivity, model Specificity and Balanced Accuracy (more appropriate validation measure than overall accuracy in the Imbalanced data), which computes the average of sensitivity and specificity based on the confusion matrix derived from test data by means of the appropriate threshold.

3 Results

3.1 Imbalanced dataset adjustment

A total of 720,612 observations were included in the dataset of which 93% of the observations were non-inundated and only 7% inundated. The same RF model ($ntree=250$, $mtry=8$) was implemented for both the imbalance and balanced dataset. Figure 3 shows the out of
bag error in the final Imbalance and Balanced RF model. The overall $OOB$ error (black line) in imbalance RF is less than in balanced RF. However, the $OOB$ error for inundated class (green dotted line) which is our target class was around 25% in the imbalance data set, conversely in the balanced data set, the error is significantly reduced to approximately 8%. Additionally, in the balanced RF model the range of error is narrow compared to the Imbalanced RF model.

**Figure 3**

In addition to $OOB$ error, the ROC curves as given in Figure 4 were also investigated. The ROC curve of Balanced RF dominates markedly the Imbalance RF model which shows more efficient performance of balanced RF. Comparison in the $AUC_{ROC}$ of Imbalance RF and balanced RF models confirms the effectiveness of the down-sampling method when used to balance data, and in this case resulted in an improved prediction of the rare class (i.e. inundated class). The average $AUC_{ROC}$ of 10-fold CV was enhanced from 0.896 for Imbalance RF to 0.915 for Balanced RF model.

**Figure 4**

### 3.2 Performance Measurement

The results of validating metrics for all the 10-fold block CV using balanced RF, are presented in Table 2. The performance metrics of Sensitivity, Specificity and Balanced Accuracy assessment were carried out according to the 0.29 probability threshold which is the average threshold of the 10 models estimated using the sensitivity-specificity sum maximizer approach. By selecting this prediction threshold instead of the default threshold of 0.50, we increase the number of true inundated observations with the cost of increasing the number of False inundated observations, although in this case it is more important for the model to predict the minority (inundated) class correctly. As such, this threshold selection approach can lead to higher model sensitivity ranging from 65 to 92%. To avoid the misleading impression of high overall accuracy given by the class imbalanced data, the balanced accuracy was used to represent the accuracy of the model, with an average accuracy of 82%.

The area under the ROC curves was used as the threshold-independent metric. $AUC_{ROC}$ is interesting as it shows how successfully the model separates the inundated and non-inundated observations and ranks them. A higher $AUC_{ROC}$ value in the test scenario signifies a better fitting model. The value of $AUC_{ROC}$ varies between 0.871 and 0.939 with the average of 0.915 which is excellent (see Bekkar et al., 2013 and Wang et al., 2015 for AUC range interpretation).
Moreover, the Standard Errors (SE) of the mean of performance metrics for the 10 models were very small, indicating that 10 runs of the individual models on new data sets (testing dataset) revealed reasonably constant prediction metrics (Table 2). This illustrates the ability of the model to perform well with an unseen dataset.

To visualize the model prediction ability, probability of inundation occurrence was mapped for the date of 2 June 1990 (which was kept out of the training dataset in fold 7 for model validation) using RF<sub>fold7</sub> model. The predicted inundation extent matched well with observation (Figure 5).

### 3.3 Predictor Importance

Two different measures are provided by RF to assess the importance of each predictor, Mean Decrease in Accuracy (MDA) and Mean Decrease in Gini (MDG). The MDG has been shown to be biased when within-predictor correlation exists (Nicodemus & Malley, 2009; Nicodemus, 2011). However, when the response class is a categorical variable with strongly unbalanced categories, MDG is expected to produce better results than MDA (Calle & Urrea 2010; Boulesteix et al., 2012). Both of these conditions exist in the data due to the nature of environmental datasets. Figure 6 shows the differences between variable importance according to MDA and MDG.

The most important predictors using MDA were ldfglobal, Q425003_30, Rainfall_20, Runoff_10, Q425003 and Runoff_30. However, this ranking was changed by the application of MDG whereas ldfglobal, Q425003_30, CostDistance, Q425003, Soilmoisture_30, and mslope were ranked as the most important predictors. To calculate the influence of each predictor on the response, the values were scaled so that the total was equal to 100 (Figure 6). The six most important variables occupy around 70% of the total contribution to inundation occurrence in both MDA and MDG. Although, the ranking of variables is slightly different using the two different measures, there are similarities which are interesting. Both of the aforementioned measures rank local deviation of topography from global (ldfglobal) and cumulative 30-day discharge at the Burke gauge Q425003_30 as the first two important predictors. Both measures rank a common suite of variables lowly: the integrated moisture index (imi); surface relief ratio (srelief); landform curvature (lform); and the compound topographic index (cti), suggesting these add little explanatory power.
4 Discussion and Conclusions

In this study, we applied a machine learning method, Random Forest, to predict daily wetland inundation at 30 m resolution using a range of predictors including hydrologic, climatic and topographic variables. This relatively high spatial resolution was required to interpret the ecological characteristics of the floodplain, on the assumption that vegetation communities recently mapped at 1:4000 for the Darling River (Florentine et al. 2014) could be understood in relation to inundation frequency, duration and inter-flood duration, variables of crucial importance to the survival and reproduction of vegetation communities (Rogers and Ralph 2010).

A monthly surface water extent at low spatial resolution (~25km grid) has recently become available covering a 15-year period (1993-2007) combining satellite observations in the visible, near-infrared and passive-active microwaves (the Global Inundation Extent from Multi-Satellites, or GIEMS: Prigent et al., 2012). Aires et al. (2017) used topographic information in a neural network inundation probability model to downscale the GIEMS data to 90m spatial resolution. They propose this as an interim approach to fine-scale global inundation modeling until the NASA-CNES Surface Water Ocean Topography (SWOT) mission will, from 2020, provide high spatial and temporal resolution of continental surface water extent and height.

These products provide widely available maps of inundation at a spatial and temporal resolution suitable for a broad range of applications. The 15-year duration of record represents a wide range of hydrological conditions and composite products may well represent inundation patterns and risks at a landscape scale. However, for our purpose, it was necessary to predict inundation on a daily time step to characterize the infrequent flooding events in the Darling River. Water propagates across the low relief floodplain over days to weeks during periods of overbank flow. The flood frequency, duration and inter-flood periodicity associated with the mosaic of floodplain vegetation requires higher temporal resolution than provided by global products, whether Landsat (Yamazaki et al., 2015; Pekel et al., 2016) or GIEMS and derivatives (Prigent et al., 2016; Aires et al., 2017). Our modelling, constrained by the Landsat archive, provides the opportunity to represent floodplain inundation associated with hydrological regimes in a daily time-step.

Further, our intention is to relate floodplain inundation to river flow at the Bourke gauge, one of a series of gauges represented in the Darling River IQQM (Integrated Quantity Quality Model), a 1-D hydrology model developed by the NSW Office of Water to represent flow in the Darling River over the full range of historical conditions. By predicting inundation on the basis of topographic, climatic and hydrological variables, we can apply a range of scenarios developed for the Darling River IQQM, including “pre-development” scenarios (hydrological modifications...
removed), and a range of climate change scenarios derived from down-scaled climatic modelling (Evans et al. 2014).

Although the number of environmental variables influencing inundation is potentially very large, including geomorphological factors, current and antecedent climatic conditions, current and antecedent river flows, the model selection procedure suggested that only a few key variables were sufficient for a satisfactory prediction, reflecting the association among the environmental variables (Siroky, 2009). The top six predictors (Figure 6) covered all the important factors influencing inundation. Water source was captured by thirty-day cumulative discharge (Q425003_30), current discharge (Q425003), cumulative rainfall of the previous 20 days (Rainfall_20) and cumulative runoff over the previous 10 and 30 days (Runoff_10 and Runoff_30). Exposure to inundation and flood retention were represented by flooding potential (CostDistance and mslope) water detention capacity (ldfglobal), and antecedent soil moisture conditions (Soilmoisture_30). Moreover, the top two predictors (i.e. ldfglobal and Q435003_30) were far more important than the others and contributed almost equally to model prediction power (Figures 6), highlighting the importance of landscape position for inundation in the semi-arid lowland floodplain (Wen et al., 2018).

The topography of the land surface (our ldfglobal) and the elevation-weighted distance to rivercourses (CostDistance) are variables similar to those utilized by Aires et al. (2017) in their downscaling of the GIEMS, and were the most important local factors contributing to flooding and water shedding. Importantly, our predictive model also incorporated as highly important two hydrological metrics modelled within the Darling River IQQM: discharge at the Bourke gauge (Q425003) and, more significantly, accumulated discharge at the said gauge in the previous 30 days. The model thereby captures the immediate and antecedent over-bank flow contributions in predicting inundation patterns across the floodplain. Similarly, the soil moisture model input, representing water within the upper meter cumulative over the previous 30 days, represents the potential losses to infiltration which can bear strongly on inundation extent depending on antecedent climatic conditions (Powell et al., 2008).

The rivers of the northern Murray Darling Basin are amongst the most hydrologically variable in the world (Puckridge et al., 1998). Floods are infrequent and occur in relation to unpredictable hydrological conditions. For this reason, most of our observations from the Landsat record are of non-inundation, and our assessment of accuracy required the balancing of inundation and non-inundation class patterns. The AUC \( \text{ROC} \) (0.896 and 0.915 for Imbalance and Balanced model, respectively) are among the “very good” range of supervised classifications (Bekkar et al., 2013; Wang et al., 2015). More importantly, the reliability of the trained models was tested using 10-fold block CV, with the results demonstrating comparable performance. For highly biased data, which is common in natural systems, results showed that the down-sampling method was an effective means by which to balance the data, and to increase the predictive power for minority classes (in this case, the inundated cell).
The computing efficiency (relative to hydrodynamic model) makes machine learning a promising approach to reproduce long-term daily inundation extent to characterize the inundation regime (i.e. timing, frequency, duration of flooding) at a floodplain scale, which is critical to compute the water requirements of various wetland ecological assets.

The most influential variables predicting inundation occurrence are topography and discharge, of which local deviation from global, 30-day accumulated discharge and elevation-weighted distance to the river ranked first, second and third. In general, geomorphological change occurs slowly (Grabowski & Gurnell, 2016) (decadal to centennial timescales).

Our model can be applied in daily time-step to compare, for example, the floodplain inundation regimes associated with current water resource development, pre-development and alternative water planning scenarios where these can be represented as flow at the Bourke gauge. It is possible also to explore the impacts of a range climate change scenarios on the ecology of the arid and semi-arid floodplains, and explore the adaptive options to mitigate the potential impacts of climate change. Such information is foundational to interpreting the implications of water sharing plans, dam operational rules and environmental flow strategies on floodplain biota at the flow event or flow regime scale. Nevertheless, one should bear in mind that the spatial applicability of the trained model is the main limitation of this approach, i.e. applying this specific fitted RF model on a completely different floodplain with dissimilar characteristics would be potentially challenging, and without further training could not be so robust. Notwithstanding, the approach can still be applied to any floodplain in the world.

Acknowledgments

Macquarie University, Australia, funded this research under the International Macquarie University Research Excellence Scholarship (iMQRES) to Sara Shaeri Karimi. Roozbeh Valavi is supported by an Australian Government Research Training Program Scholarship and a Rowden White Scholarship. The data sources used in this study are cited and explained in the text.

References


Álvarez-Cabria, M., Barquin, J., & Peñas, F. J. (2016). Modelling the spatial and seasonal variability of water quality for entire river networks: Relationships with natural and


dryland vegetation response to flooding and drought at sub-continental scale. Remote


criteria for wetlands along the Barwon-Darling River. CRC for Freshwater Ecology


Threats to Intermittent Rivers and Ephemeral Streams. In T. Datry, N. Bonada, & A.
Boulton (Eds.), Intermittent Rivers and Ephemeral Streams (pp. 433-454): Academic

University of California, Berkeley, 110, 1-12.


https://doi.org/10.1890/07-0539.1.

D. Rosa, I. M., Marques, A. T., Palminha, G., Costa, H., Mascarenhas, M., Fonseca, C., &
Bernardino, J. (2016). Classification success of six machine learning algorithms in radar


Mueller, N., Lewis, A., Roberts, D., Ring, S., Melrose, R., Sixsmith, J., et al. (2016). Water observations from space: Mapping surface water from 25 years of Landsat imagery across...


This article is protected by copyright. All rights reserved.
## Tables

**Table 1. List of Independent Variables of the Model**

<table>
<thead>
<tr>
<th>Independent variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Topographic parameters</strong></td>
<td></td>
</tr>
<tr>
<td>cti</td>
<td>Compound Topographic Index: A steady state wetness index (Gessler et al., 1995)</td>
</tr>
<tr>
<td>lform</td>
<td>Landform Curvature: Surface Curvature (concavity/convexity) index (McNab, 1993)</td>
</tr>
<tr>
<td>ldfglobal</td>
<td>Local deviation from global: Indicates the local (cell) deviation from a specified global (window) statistic (Tesfa &amp; Leung, 2017). In this study, we used the mean of a 3×3 window.</td>
</tr>
<tr>
<td>imi</td>
<td>Integrated Moisture Index: An estimate of soil moisture in topographically heterogeneous landscapes (Iverson et al., 1997)</td>
</tr>
<tr>
<td>srelief</td>
<td>Surface Relief Ratio: describes rugosity in a continuous raster surface within a specified window (Pike, 1971)</td>
</tr>
<tr>
<td>mslope</td>
<td>Mean of slope within a 3 m ×3 m window</td>
</tr>
<tr>
<td>CostDistance</td>
<td>the Euclidean distance to streams and rivers weighted by elevation</td>
</tr>
<tr>
<td><strong>Climatic parameters</strong></td>
<td></td>
</tr>
<tr>
<td>Rainfall_5</td>
<td>Accumulated rainfall value of prior 5 days</td>
</tr>
<tr>
<td>Rainfall_20</td>
<td>Accumulated rainfall value of prior 20 days</td>
</tr>
<tr>
<td>Runoff</td>
<td>Daily runoff</td>
</tr>
<tr>
<td>Runoff_10</td>
<td>Accumulated runoff value of prior 10 days</td>
</tr>
<tr>
<td>Runoff_30</td>
<td>Accumulated runoff value of prior 30 days</td>
</tr>
<tr>
<td>Soilmoisture_30</td>
<td>Root Zone Soil Moisture represents the percentage of available water content in the top 1 m of the soil profile (cumulative value of prior 30 days)</td>
</tr>
<tr>
<td><strong>Hydrologic parameters</strong></td>
<td></td>
</tr>
<tr>
<td>Q425003</td>
<td>Daily discharge at Bourke station</td>
</tr>
<tr>
<td>Q425003_30</td>
<td>Accumulated discharge value of prior 30 days at Bourke station</td>
</tr>
</tbody>
</table>
Table 2. Performance Metrics of 10-fold Block Cross-Validation Using Test Dataset

<table>
<thead>
<tr>
<th>Fold</th>
<th>Sensitivity (%)</th>
<th>Specificity (%)</th>
<th>Balanced Accuracy (%)</th>
<th>AUC&lt;sub&gt;ROC&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>65.6</td>
<td>89.9</td>
<td>77.7</td>
<td>0.893</td>
</tr>
<tr>
<td>2</td>
<td>88.9</td>
<td>84.5</td>
<td>86.7</td>
<td>0.932</td>
</tr>
<tr>
<td>3</td>
<td>80.1</td>
<td>86.0</td>
<td>83.0</td>
<td>0.926</td>
</tr>
<tr>
<td>4</td>
<td>81.3</td>
<td>82.9</td>
<td>82.1</td>
<td>0.916</td>
</tr>
<tr>
<td>5</td>
<td>84.8</td>
<td>65.3</td>
<td>75.0</td>
<td>0.871</td>
</tr>
<tr>
<td>6</td>
<td>87.0</td>
<td>80.2</td>
<td>83.6</td>
<td>0.929</td>
</tr>
<tr>
<td>7</td>
<td>81.2</td>
<td>86.4</td>
<td>83.8</td>
<td>0.907</td>
</tr>
<tr>
<td>8</td>
<td>67.7</td>
<td>93.6</td>
<td>80.6</td>
<td>0.903</td>
</tr>
<tr>
<td>9</td>
<td>92.0</td>
<td>80.0</td>
<td>86.0</td>
<td>0.939</td>
</tr>
<tr>
<td>10</td>
<td>94.2</td>
<td>74.3</td>
<td>84.2</td>
<td>0.929</td>
</tr>
<tr>
<td>Average</td>
<td>82.3</td>
<td>82.3</td>
<td>82.3</td>
<td>0.915</td>
</tr>
<tr>
<td>SE</td>
<td>3.0</td>
<td>2.6</td>
<td>1.1</td>
<td>0.007</td>
</tr>
</tbody>
</table>
Figure captions

Figure 1. Location of the study area within the Murray-Darling Basin, NSW, Australia, including digital elevation model (DEM) of the floodplain.

Figure 2. RF parameters tuning: (a) mtry vs OOB error and (b) ntree vs OOB error.

Figure 3. RF OOB Error for Imbalance (a) and Balanced model (b). Green dotted line is Inundated class, Black line is overall error and the red dashed line is non-inundated class.

Figure 4. Comparison of ROC curves for Balanced and Imbalance RF.

Figure 5. Probabilities of inundation occurrence generated through balanced RFfold7 for 2 June 1990 (a). Actual inundation (2 June 1990) left out of the training set in fold 7 for validation (b).

Figure 6. Comparison of MDA and MDG Variable importance.
Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Karimi, SS;Saintilan, N;Wen, L;Valavi, R

Title:
Application of Machine Learning to Model Wetland Inundation Patterns Across a Large Semiarid Floodplain

Date:
2019-11

Citation:

Persistent Link:
http://hdl.handle.net/11343/286590