Environmental Modelling and Software 135 (2021) 104907

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: http://www.elsevier.com/locate/envsoft

A wavelet-based tool to modulate variance in predictors: An application to predicting drought anomalies

Ze Jiang [a], Md. Mamunur Rashid [b], Fiona Johnson [a], Ashish Sharma [a,*]

[a] School of Civil and Environmental Engineering, University of New South Wales, Sydney, New South Wales, Australia
[b] Civil, Environmental, and Construction Engineering & National Center for Integrated Coastal Research, University of Central Florida, Orlando, FL, USA

ARTICLE INFO

Keywords:
Wavelet system prediction
R package
Maximal overlap discrete wavelet transform
Unbiased estimator

ABSTRACT

This work presents an open-source tool to predict natural system responses by transforming the frequency spectrum of predictor variables to create a response that better resembles observations. The R package, namely WAvelet System Prediction (WASP), is based on a discrete wavelet transform (DWT)-based variance transformation method. We further introduce the maximal overlap DWT (MODWT)-based variance transformation which allows the method to be used in forecasting applications. We also develop the method to include an unbiased estimator that mitigates the well-known issue of edge effects in wavelet transforms. The predictive model in the method is a k-nearest neighbor (knn) approach. The main functionalities of the software include: (1) transforming the system predictors, (2) identifying significant predictors corresponding to the response, (3) predicting target response using the knn. Results of predicting sustained drought anomalies across Australia show clear improvements in predictive skill compared to the use of untransformed predictors.

1. Introduction

A regression model describes the relationship between a system response and a finite set of predictor variables using an assumed modelling form (linear or nonlinear). Approaches can range from simple regression models using a range of physically justifiable predictor variables (Hertig and Tramblay, 2017) to those where more complex transformations including rotations are adopted (Jiang et al., 2019; Ndehedehe et al., 2016). Differing spectral attributes in a response and a system predictor can complicate specifying the system-prediction model. We present here an approach that addresses this difficulty by optimally transforming each predictor variable to better characterise the spectrum of the response being modelled. The underlying idea behind the approach here is to improve the modelling of natural systems where the potential predictor variables vary at time scales that differ from those of the plausible response. For example, in hydrology, daily precipitation is used to predict catchment streamflow. However, attenuation from catchment storage means that at short time-scales the variability of streamflow is substantially dampened compared to rainfall. Thus conventional regression modelling approaches can have difficulties in characterising this differing variability and formulating a relationship (Rashid and Beecham, 2019). Although the approach is generic and can be used for any natural system model, our specific focus is on hydro-climatological systems. An example of such a prediction problem is the need to assess changes to natural systems due to climate change. In this case, Earth System Models and/or General Circulation Models (GCMs) provide predictors that can be used to model future changes in hydrological variables.

Alternatives to transfer the modelling problem into the frequency domain include methods such as Fourier and wavelet transforms. Our approach uses wavelet theory to formulate an optimal model of the system to improve the assessment of changes into the future. The wavelet transform (WT) is adopted in the approach to avoid loss of temporal information when transforming to the frequency domain using a Fourier transform (Daubechies, 1990; Strang, 1996; Torrence and Compo, 1998). The WT can decompose the original time series into separate large-scale (slowly changing, low frequency) and fine-scale (rapidly changing details, high frequency) time series. A number of models based on frequency domain analysis have been proposed recently to simulate and predict the variability in the response (Fahimi, Yaseen, & El-shafie, 2017; Nguyen et al., 2019; Quilty et al., 2019; Rashid et al., 2018; Sang, 2013). Most of those applications directly use the decomposed time series to forecast the target response (Quilty and Adamowski, 2018; Rashid et al., 2016). Jiang, Sharma, and Johnson

* Corresponding author.
E-mail address: a.sharma@unsw.edu.au (A. Sharma).

https://doi.org/10.1016/j.envsoft.2020.104907
Accepted 14 October 2020
Available online 17 October 2020

(2020) proposed a new approach by using the decomposed time series to reconstruct a new set of predictor variables that can explain maximal information in the response. They showed that this approach can significantly improve the performance of the regression model, when applied firstly to synthetic data and a drought index downscaling case study at fifteen rainfall gauges in Sydney, Australia. However, the original method is limited to prediction problems where the future state of the predictors is "known", which is due to the mathematical properties of the discrete wavelet transform (DWT). If a forecasting model is required then DWT is not suitable because this wavelet transform requires future information (which is not available in a forecasting setting) to predict the target response (Du et al., 2017; Quilty and Adamowski, 2018). To address this issue, other wavelet transformations can be used to implement the variance transformation, including the maximal overlap DWT (MODWT) and à trous algorithm (AT). In this case, DWT can be replaced with MODWT or AT, which have no dependence on future information (Nason, 2008; Quilty and Adamowski, 2018). Therefore, we have included the MODWT and AT based variance transformation into the WASP R-package. Alternatively, when considering climate change projections, the DWT forecasting problem is overcome because reliable future projections of the predictors are available from GCMs although the target response is unknown or its projection is not reliable as shown by Rashid et al. (2018) and Fowler et al. (2007).

Another issue in using wavelet-based methods in real-world applications is the edge effects resulting mainly due to limited sample sizes, also known as the error due to the boundary condition that is associated with wavelet decompositions, including wavelet and scaling coefficients (Percival and Walden, 2000). However, there are ways to reduce the effects of boundary bias in wavelet transformations. An estimator excluding the boundary coefficients is regarded as an unbiased wavelet variance estimator (Cornish et al., 2006). This logic can be applied to the proposed variance transformation method, which leads to an unbiased variance transformation. Thus, the methodological contributions of this study are to: (1) generalise the wavelet-based variance transformation method to allow it to be applied in forecasting problems and (2) develop an unbiased variance transformation. This substantially broadens the application of the proposed method across a wide range of systems beyond the simplified illustrations in Jiang et al. (2020).

The approach outlined above is embodied in the WAvelet System Prediction (WASP) R-package, and it consists of three key functions. The first function finds the optimal variance transformation for each predictor variable of interest, reconstructing a new predictor with complimentary spectral attributes to the predictand. The second function identifies significant reconstructed predictors using partial informational correlation (PIC). PIC is used to measure the dependence between a given response and the reconstructed new predictor conditioned to pre-existing predictor(s) (Sharma, 2000; Sharma and Mehrotra, 2014). The last function is the predictive model, which is a *k*-nearest neighbor (knn) estimator using a kernel regression function (Lall and Sharma, 1996; Sharma et al., 1997). An additional contribution here is that the knn estimator has been modified to better allow for extrapolation.

In this study, we implement the MODWT-based and unbiased variance transformation in the R-package WASP and evaluate it on a large scale of hydro-climatological system. For instance, sustained droughts are natural hazards associated with a range of climatic factors such as low precipitation and high temperatures and potential evapotranspiration (Sheffield, 2011). These climatic factors are in turn affected by large scale climate teleconnections which vary over periods of years to decades (e.g., El Niño Southern Oscillation and India Ocean Dipole) (Mishra and Singh, 2010), as well as long-term trends from anthropogenic climate change (Dai, 2013; Sheffield and Wood, 2008). Thus, drought is a result of the interactions of a large number of variables all of which have very different spectral properties. Here the variance transformation method is demonstrated by modelling and predicting sustained drought anomalies for Australia as represented by the Standardized Precipitation Index (SPI).

2. Methodology

2.1. MODWT-based variance transformation

In this section, we first introduce the original DWT-based variance transformation and then extend it to include the MODWT-based variance transformation. Full details and derivation of the variance transformation are provided in Jiang et al. (2020). A summary of the important steps is provided here. Consider a set of n paired centred (i.e., with mean of zero) observations of the predictor variable $\boldsymbol{X}$ and the response variable $\boldsymbol{Y}$, i.e., $(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})$. First, the signal $\boldsymbol{X}$ is decomposed into a vector of coefficients matrix $\mathbf{W} = [\mathbf{D}_1, \ldots, \mathbf{D}_J, \mathbf{A}_J]$ with a dimension of $n \times 1$ using the DWT. The coefficients matrix is then reconstructed into a matrix of frequency components $\mathbf{R} = [\mathbf{d}_1, \ldots, \mathbf{d}_J, \mathbf{a}_J]$, and the associated variance structure of these sub-time series is given by $\mathbf{I} = [\sigma_{d_1}, \ldots, \sigma_{d_J}, \sigma_{a_J}]^T$ (Percival and Walden, 2000). This is so-called multiresolution analysis (MRA). Here, J is the highest decomposition level, which will be further discussed in the section of unbiased variance transformation. The property of DWT ensures that the sum of the variance of the sub-time series equals the variance of the original time series, which means $\sum_{j=1}^{J+1} \mathbf{I}_j^2 = \frac{1}{n-1}\mathbf{X}^{\mathbf{T}}\mathbf{X} = \sigma_{\mathbf{X}}^2$. Accordingly, $\boldsymbol{X}$ can be written as a matrix multiplication $\mathbf{X} = \widehat{\mathbf{R}}\mathbf{I}$ with the standardized reconstruction matrix $\widehat{\mathbf{R}} = [\widehat{\mathbf{d}}_1, \ldots, \widehat{\mathbf{d}}_J, \widehat{\mathbf{a}}_J]$. The variance transformation is achieved by reconstructing a new predictor $\boldsymbol{X}'$ with variance structure $\boldsymbol{\alpha}$ similar to the corresponding response in the frequency domain. They can be written as:

$$\begin{aligned} \mathbf{X}' &= \widehat{\mathbf{R}}\boldsymbol{\alpha} \\ \boldsymbol{\alpha} &= \sigma_X \widehat{\mathbf{C}} \end{aligned} \tag{1}$$

where $\widehat{\mathbf{C}}$ is the normalized covariance of the variable set $(\mathbf{Y}, \widehat{\mathbf{R}})$, and the covariance $\boldsymbol{C}$ has the form of

$$\mathbf{C} = \frac{1}{n-1}\mathbf{Y}^T\widehat{\mathbf{R}} = \left[S_{Y\widehat{d}_1}, \ldots, S_{Y\widehat{d}_J}, S_{Y\widehat{a}_J}\right]. \tag{2}$$

Essentially, the reconstructed new predictor $\boldsymbol{X}'$ is obtained by redistributing the variance in its spectrum and it has the same total variance as the original predictor $\boldsymbol{X}$. All potential predictors will be reconstructed with this operation, and a reconstructed new set of predictors is then used for predictor selection and response prediction. Assuming that the variance transformed predictor is used to predict the associated response with simple linear regression, we can derive the theoretical optimal prediction accuracy as measured by root mean square error (RMSE):

$$RMSE_{\min} = \sqrt{\frac{n-1}{n}\left(\sigma_Y^2 - \|\mathbf{C}\|^2\right)}, \tag{3}$$

where σ_Y denotes the standard deviation of the response Y.

The method originally proposed by Jiang et al. (2020) requires both additive decomposition (i.e., MRA) and variance decomposition (i.e., energy-based decomposition). To extend the method to consider forecasting problems, the DWT can be replaced by wavelet approaches that do not include future time steps (such as MODWT and AT). However, for the above derivation to be valid then the new wavelet approaches need to also fulfill the requirement for additive and variance decomposition. Both MODWT and AT fulfill these two requirements only when the Haar wavelet filter (equivalent to Daubechies 1, db1 or d2) is adopted. When the Haar wavelet filter is used MODWT and AT are equivalent (i.e., lead to the same decomposed frequency components). Therefore, for forecasting applications, WASP has been extended to include MODWT with the Haar wavelet filter as the basis for the variance transformation. There is a potential risk that the spectrum of the variables of interest cannot be characterised well because the wavelet filter is limited to the

Table 1
Summary of the size of boundary effects.

Wavelet Method	Beginning of the signal	End of the signal	Total number of boundary coefficients	Non-boundary coefficients
DWT-MRA	t = 0,1, …, L_j-2	t = N-1, N-2, …, N-L_j+1	2(L_j-1)	N-2(L_j-1)
MODWT	t = 0,1, …, L_j-2	–	L_j-1	N-L_j+1

Note: The width of the j-th level wavelet or scaling filter $L_j = (2^j\text{-}1)(L\text{-}1)+1$, where L is the width of the $j = 1$ base filter.

Haar wavelet filter. However, the logic can be applied to both MODWT and AT when other wavelet filters are adopted, as they provide additional ways to characterise the spectrum of variables of interest.

Another advantage of using MODWT is that there is no restriction on the dyadic sample size. Briefly, MODWT decomposes the original time series $\boldsymbol{X}$ into a $n \times (J+1)$ matrix of wavelet and scaling coefficients $\tilde{\mathbf{W}} = [\tilde{\mathbf{D}}_1, \ldots, \tilde{\mathbf{D}}_J, \tilde{\mathbf{A}}_J]$, and the associated standard deviation matrix is given by $\tilde{\mathbf{I}} = [\sigma_{\tilde{D}_1}, \ldots, \sigma_{\tilde{D}_J}, \sigma_{\tilde{A}_J}]^T$. MODWT also ensures variance decomposition, which means $\sum_{j=1}^{J+1} \tilde{\mathbf{I}}_j^2 = \frac{1}{n-1}\mathbf{X^T X} = \sigma_{\mathbf{X}}^2$. This provides a way to investigate and transform the variance structure of the coefficients matrix, $\tilde{\mathbf{W}}$, directly. As a result, using the covariance $\boldsymbol{C}$ of the variable set $(\mathbf{Y}, \tilde{W})$ the variance transformed $\boldsymbol{X}'$ can be obtained given by the equation:

$$\mathbf{X}^{'} = \tilde{\mathbf{W}}\boldsymbol{\alpha}\widehat{\boldsymbol{\alpha}} = \sigma_X \widehat{\mathbf{C}} \tag{4}$$

where $\widehat{\tilde{\mathbf{W}}}$ is the standardized coefficients matrix $\tilde{\mathbf{W}}$. It is noted that the coefficients matrix $\mathbf{W}$ decomposed from DWT has the dimension of $n \times 1$ while the coefficients matrix $\tilde{\mathbf{W}}$ from MODWT has a dimension of $n \times (J+1)$. Expect for the independence on future information, this is another reason the coefficients matrix of MODWT can be directly used for variance transformation.

2.2. Unbiased variance transformation

The second methodological contribution of this study is to solve the issue of boundary bias in applying wavelet-based methods. Boundary related issues are due to sample size, the choice of decomposition level, as well as wavelet filter. Table 1 summarizes the size of the boundary effects for both types of wavelet transforms. As shown in the table, the number of non-boundary coefficients depends on the sample size (N), the decomposition level (j), and the width of wavelet filter (L). The multiresolution analysis of DWT is affected at the beginning and the end of the sub-time series while MODWT is only affected at the start of the decomposed components. It is clear that shorter wavelet filter, longer time series or lower decomposition level leads to a smaller number of boundary coefficients. In wavelet theory, the exclusion of boundary coefficients in wavelet variance estimation is called unbiased estimator (Cornish et al., 2006). There is a smaller difference between biased and unbiased estimates when fewer boundary coefficients need to be excluded.

Here we propose to adopt the unbiased variance transformation by computing the covariance using only the non-boundary coefficients as follows:

$$\mathbf{C}^* = \frac{1}{n-1}\mathbf{Y}^T\widehat{\mathbf{R}}^* \tag{5}$$

where the asterisk * implies the unbiased value. $\widehat{\mathbf{R}}^*$ (or $\widehat{\tilde{W}^*}$) is the standardized matrix excluding boundary coefficients, and $\mathbf{C}^*$ is a vector of unbiased covariance. It is worth noting that the unbiased estimator can only be computed for some decomposition levels. However, the nature of variance transformation requires greater decomposition levels thus we still use the biased estimator whenever the unbiased estimator is not available. The introduction of unbiased variance transformation is not likely to change the model performance substantially when a shorter wavelet filter is used and larger sample size is available.

2.3. Partial informational correlation

The wavelet-based variance transformation approach adopts PIC, which takes the partial dependence between predictors and the response into account to identify significant (in this case variance transformed) predictors. A short description of the logic behind PIC is presented here, and readers are referred to Sharma (2000), Galelli et al. (2014) and Sharma and Mehrotra (2014) for additional details, as well as to Sharma et al. (2016) for the software, known as NPRED, needed to estimate the PIC.

The partial information (PI) is based on information theory and measures the dependence between the response Y and a potential predictor X of the response given pre-existing predictor(s) $\boldsymbol{Z}$. Thus, a sample estimate of $PI(Y, X|\mathbf{Z})$ is written as:

$$\widehat{PI}(Y,\ X \mid \mathbf{Z}) = \frac{1}{n}\sum_{i=1}^{n} \log_e \left[\frac{f_{Y|\mathbf{Z},X|\mathbf{Z}}(Y_i, X_i|\ \mathbf{Z_i})}{f_{Y|\mathbf{Z}}(Y_i|\mathbf{Z_i}) f_{X|\mathbf{Z}}(X_i \mid \mathbf{Z_i})} \right] \tag{6}$$

where Y_i and X_i is the i-th bivariate sample data pair in a sample of size n. $Y|\mathbf{Z}$ and $X|\mathbf{Z}$ are partial response and partial independent variable, which represent the residual information in variables Y and X, when the effect of pre-existing predictor(s) $\boldsymbol{Z}$ has been taken into account. $f_{Y|\mathbf{Z}}(Y_i|\mathbf{Z}), f_{X|\mathbf{Z}}(X_i|\mathbf{Z})$ and $f_{Y|\mathbf{Z},X|\mathbf{Z}}(Y_i, X_i|\mathbf{Z})$ are the respective marginal and joint probability densities using kernel density estimation. The PI can be scaled to the range from 0 to 1, which is introduced as the PIC:

$$\widehat{PIC} = \sqrt{1 - \exp\left(-2\widehat{PI}\right)} \tag{7}$$

Thus, the PIC is a generic measure of conditional dependence, where 0 represents no dependence and 1 represents perfect dependence. A measure of statistical significance for the PIC is also required,

$$t = PIC\sqrt{\frac{m}{1 - PIC^2}} \tag{8}$$

where t follows the Student's t distribution with $m = n\text{-}l$ degrees of freedom, with n being the number of observations and l the number of conditioning variables. This is used for the stopping criterion when selecting the significant predictor variable(s). Given a certain significance level p (we used $p = 0.1$ in the case study), when the estimated PIC is smaller than an associated threshold PIC_p for all the remaining partial predictors, the selection process will be terminated.

2.4. Modified k-nearest neighbor regression estimator

Selecting a predictive model is generally based on the nature of the modelling system as well as the modeler's experience. Regression methods have been widely solved by using the parametric least squares estimator approach. Non-parametric models can also be used with the advantage that fewer assumptions about the distribution of the population are required. In this study, the nonparametric knn method was used for prediction.

The key to the knn method is to find the closest observations to x in the training dataset to form $\widehat{Y}$. Specifically, the knn fit for $\widehat{Y}$ is defined as follows:

$$\widehat{Y}(x) = \frac{1}{k}\sum_{x_i \in N_k(x)} y_i \tag{9}$$

where $N_k(x)$ is the neighbourhood of x defined by the k closest points x_i

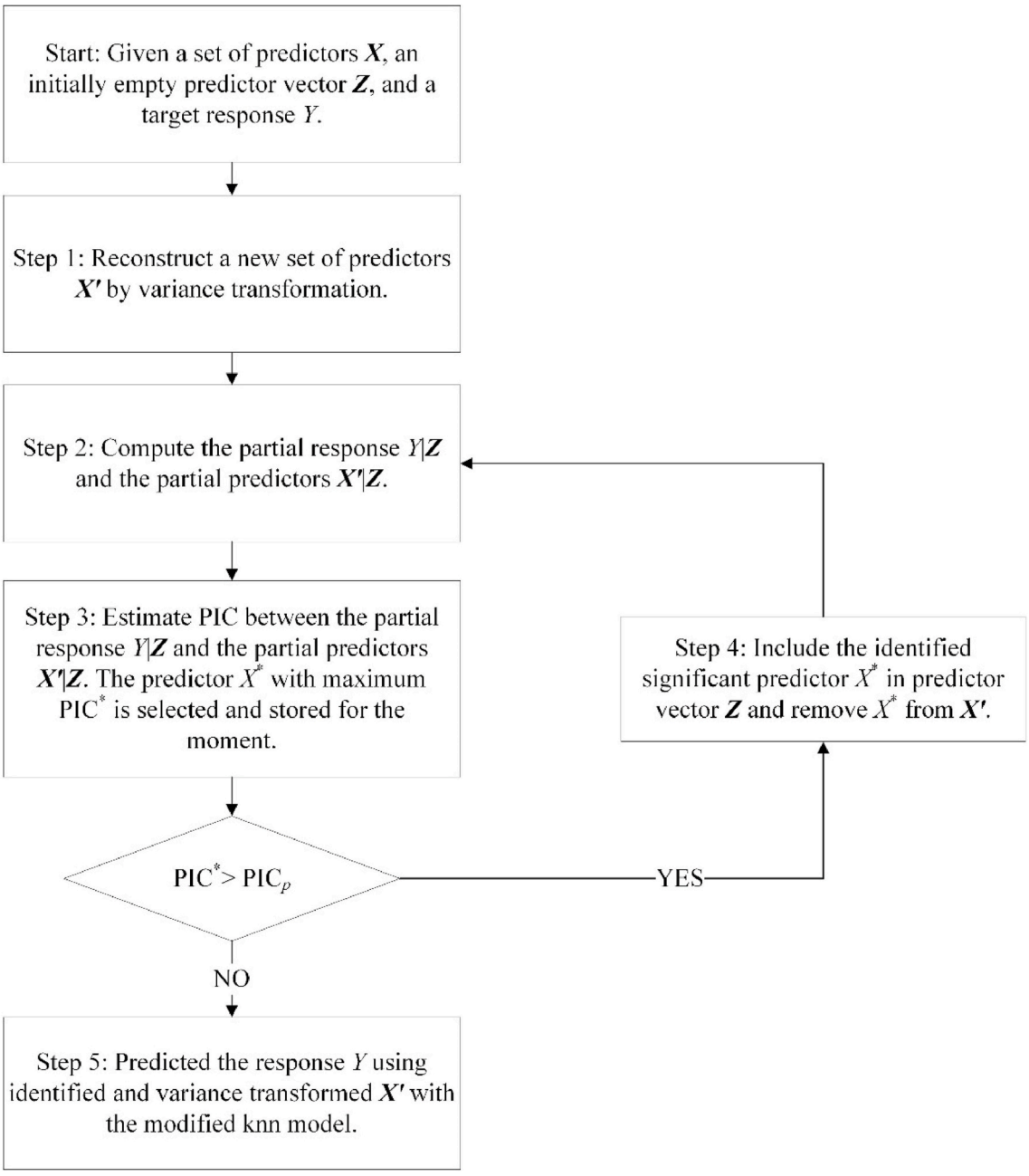

Fig. 1. Flowchart of the proposed variance transformation method.

in the training sample (Friedman et al., 2001). The closeness is a distance metric, which can be defined by Euclidean distance, as well as a range of alternate distance metrics (Weinberger et al., 2006). Mehrotra and Sharma (2006a) argue for the use of a weighted Euclidean distance using a discrete kernel, $K(i)$ with weights estimated based on the partial importance each predictor exerts on the response. In this current study, a linear extrapolation of the associated response based on the covariance of the predictor-response dataset was implemented. This is required because when considering climate change projections future predictor values may exceed the range of the observed data over which the knn model is trained. This subtle modification is based on the kernel regression as described by Sharma et al. (1997):

$$\hat{Y}\left(x\right) = \frac{1}{K(i)} \sum_{x_i \in N_k(x)} \left(y_i + S_{xy}^T S_{xx}^{-1} \left(x_i - x\right)\right)$$
$$K\left(i\right) = \frac{1/i}{\sum_{i=1}^{k} 1/i} \tag{10}$$

where S_{xy} and S_{xx} represent the covariance matrix for the variable set (x, y) and (x, x), respectively.

3. WASP R-package structure

3.1. Details of the software

Fig. 1 is a flowchart of the proposed method, showing the general

```
#-------------------------------------------------------------------------------
#load response, predictors variables, and grid index
data(SPI.12); data(data.CI); data(Ind_AWAP.2.5)
#study grids and period
Grid <- sample(Ind_AWAP.2.5,1)
SPI.12.ts <- window(SPI.12, start=c(1910,1),end=c(2009,12))
data.CI.ts <- window(data.CI, start=c(1910,1),end=c(2009,12))
#partition into two folds
folds <- cut(seq(1,nrow(SPI.12.ts)),breaks=2,labels=FALSE)
sub.cali <- which(folds==1, arr.ind=TRUE); sub.vali <- which(folds==2, arr.ind=TRUE)
#-------------------------------------------------------------------------------
###calibration and selection
data <- list(x=SPI.12.ts[sub.cali,Grid],dp=data.CI.ts[sub.cali,])

#variance transformation - calibration
modwt <- modwt.vt(data, wf="haar", J=8, boundary="periodic")

#stepwise PIC selection
sel <- NPRED::stepwise.PIC(modwt$x, modwt$dp.n)
#-------------------------------------------------------------------------------
###validation and prediction
data.val <- list(x=SPI.12.ts[sub.vali,Grid],dp=data.CI.ts[sub.vali,])

#variance transformation - validation
modwt.val <- modwt.vt.val(data.val, J=8, modwt)

#knn prediction
cpy <- sel$cpy; wt <- sel$wt
x=data$x; z=modwt$dp.n[,cpy]; zout=modwt.val$dp.n[,cpy]
mod <- knn(x, z, zout, k=5, pw=wt, extrap=T)
```

Fig. 2. Example of typical usage of modwt.vt and modwt.vt.val for the real case study at a sampled grid. Here the task is to transform potential predictors (climate indices), identify the significant predictors, and predict the associated response (SPI12) using a modified knn model. Note that the predictor selection uses the stepwisePIC function directly from the NPRED package.

process that is required for the variance transformation technique. This algorithm is implemented in the R library WASP software. A detailed help-file for each function and test data are provided in the package as well.

In summary, the R package consists of built-in functions for variance transformation operation for calibration ("dwt.vt", "modwt.vt", and "at.vt") and validation ("dwt.vt.val", "modwt.vt.val", and "at.vt.val") based on DWT, MODWT, and AT, respectively; the option of unbiased variance transformation for each variance transformation method is included in these functions with flag = c("biased", "unbiased"); and the modified knn regression predictive model ("knn"). There are several supplementary functions, including padding function ("padding") which extends the data to provide a dyadic sample size for the DWT-based variance transformation, and three synthetic data generator functions used in Jiang et al. (2020). Each of these codes come with associated help-files that provide guidance on their use. As described in the following section, datasets from the drought prediction case study are provided in the package, and all the results reported in this paper are reproducible using RMarkdown provided in the vignettes of this R package. Fig. 2 is a screenshot of the sequence of R commands illustrating the usage of the WASP package to transform the potential predictors (see Figure S 1 in the Supporting Material for an example of predictor variables before and after variance transformation corresponding to the response), identify the significant predictors, and predict the associated response. MODWT is adopted as the basis of wavelet transform in this case study since we are using observed data to predict target response and thus there is no dependence on future information. All codes and data in the package are open source.

It should be noted that when applying this method to forecast a future response, we assume that the conditional dependence between the predictor variables and the response remains unchanged into the future. Thus, the covariance between the response and wavelet decompositions of predictor variables from historical data is used for future predictions as well as the fitted predictive model. To check the validity of this assumption, we use cross-validation by partitioning the historical data into four complementary subsets. One subset is used as the validation set while other subsets are used as the calibration set. The results presented hereafter are cross-validated results for the entire period. The rationale for using cross-validation is that we can have a better assessment of the model performance with independent datasets (Mehrotra and Sharma, 2006b; Nguyen et al., 2019). It is important to note however that in the context of anthropogenic climate change, the range of future changes will likely exceed those observed in the past so the cross-validation is not a perfect test of our stationarity assumption for the predictor-response dependence structure.

3.2. Prediction of Standardized Precipitation Index over Australia

The WASP package was applied to predict the SPI using various climate indicators over Australia to assess the impact the variance transformation makes. We adopted climate indicators used in previous prediction of sustained hydrologic anomalies using the SPI (Rashid et al., 2020) and further expanded this dataset by including additional climate drivers strongly influencing Australia climate (Cai and Cowan,

Table 2
List of atmospheric variables considered in the study.

Index No.	Climate Indicator	Abbreviation
1	East Central Tropical Pacific, the area averaged SST from 5S-5N and 170-120W.	Nino3.4
2	Pacific Decadal Oscillation, the leading PC of monthly SST anomalies in the North Pacific Ocean	PDO
3	Southern Annular Mode, the difference of zonal mean SLP between 40°S and 65°S	SAM
4	Indian Ocean Dipole, the anomalous SST gradient between the western equatorial Indian Ocean (50E-70E and 10S-10N) and the south eastern equatorial Indian Ocean (90E-110E and 10S-0N), named as Dipole Mode Index	DMI

2013; Kirono et al., 2010; Murphy and Timbal, 2008). Table 2 lists the details of the climate indices considered in this study. The monthly anomalies of Nino3.4, PDO, and DMI are derived from monthly sea surface temperature (SST) values of Hadley Centre Global Ice and Sea Surface Temperature (HadISST) datasets (Rayner et al., 2003). SAM is calculated using sea level pressure (SLP) from NOAA Earth System Research Laboratory's Physical Sciences Division (PSD). The Australian Water Availability Project (AWAP) gridded monthly rainfall metadata is obtained from the Bureau of Meteorology (Jones et al., 2009) and is regarded as observations. The rainfall data was re-gridded to 2.5 ° × 2.5 ° over Australia using weighted area average and the SPI for 12-month and 36-month periods (SPI12/SPI36) is calculated (McKee et al., 1993). Note that grid cells where more than 25% of rainfall values are zero or missing are removed from the calculations due to data reliability concerns (Spinoni et al., 2014). As described previously, we split the data into four equal subsets for cross-validation. The study period was 1910–2009.

First of all, significant climate indicators were identified at each rainfall grid over Australia using PIC from the set of four variance transformed climate indices. In Fig. 3, the most significant drivers (i.e., the most frequently selected predictor in the PIC process among the four cross-validation subsets) for both SPI12 and SPI36 are shown. In addition, four randomly chosen grids that are used to examine the results in detail in this study are highlighted with grid index numbers in red color (refer to Figure S 2 in the Supporting Material for the complete overview of grid index over Australia). Table 3 summarizes the most significant climatic driver selection using both original and variance transformed (VT) climate indices (see Figure S 3 in the Supporting Material for the selection results using original climate indices).

As expected based on previous research, drought in Australia is significantly influenced by El Niño–Southern Oscillation (ENSO) (Cai et al., 2011; Fierro and Leslie, 2013; Pui et al., 2012; Westra and Sharma, 2010). For SPI12, most grids (83%) are influenced by ENSO, and more than half of grids (89%) are sensitive to ENSO for SPI36. On the other hand, the selection results using original climate indices are similar (i.e., ENSO is the main climatic driver of Australia rainfall) with less grids affected by ENSO particularly for SPI36. One interesting observation is that there are several grids where no climate drivers are identified as useful for prediction if the original (untransformed) climate indices are used because of the discrepancies in the temporal scale of the response and the potential predictors. This demonstrates the advantage of variance transformation technique in selecting predictor variables (Jiang et al., 2020). Another interesting outcome is that the use of variance transformation leads to a reduced selection of non-Pacific variability indicators such as DMI and SAM in the resulting model, with these variables being relegated to second or higher order predictors in the ensuing model.

Fig. 4 (a) and (c) present the density plots of observed, predicted and predicted with variance transformation SPI at the four randomly sampled grids. It is clear that the probability distributions of predicted SPI using variance transformed predictors are closer to observed SPI in the sampled grids. Its closeness can also be measured by the PDF skill scores (Perkins et al., 2007), which are shown in Fig. 4 (b) and (d). The value of a PDF skill score ranges between 0 and 1, and 1 represents a perfect match. These results suggest that the wavelet-based approach

Table 3
Number of grid cells with significant order 1 predictor variable of SPI over Australia with and without variance transformation.

Drought Index	Model	Nino34	PDO	SAM	DMI	Total (138)
SPI12	VT	114	14	7	3	138
SPI36		123	11	4	0	138
SPI12	Original	98	13	12	13	136
SPI36		63	38	22	12	135

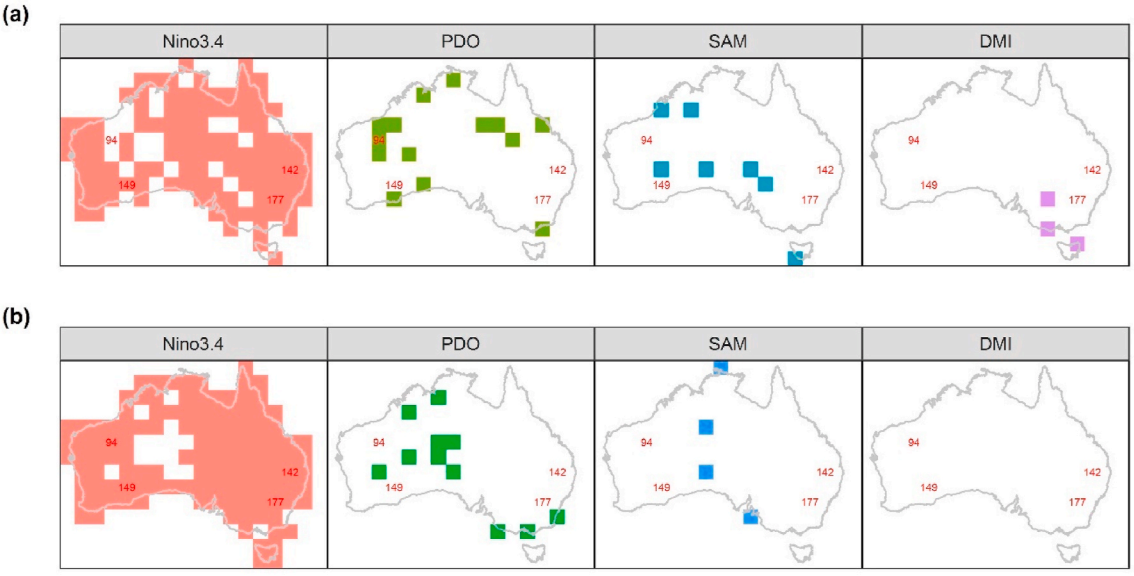

Fig. 3. The most significant predictors identified using variance transformed climate indices over Australia for different time scales of SPI. **(a) SPI12**; **(b) SPI36**. Four randomly sampled grids investigated in the study are indicated in red color. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

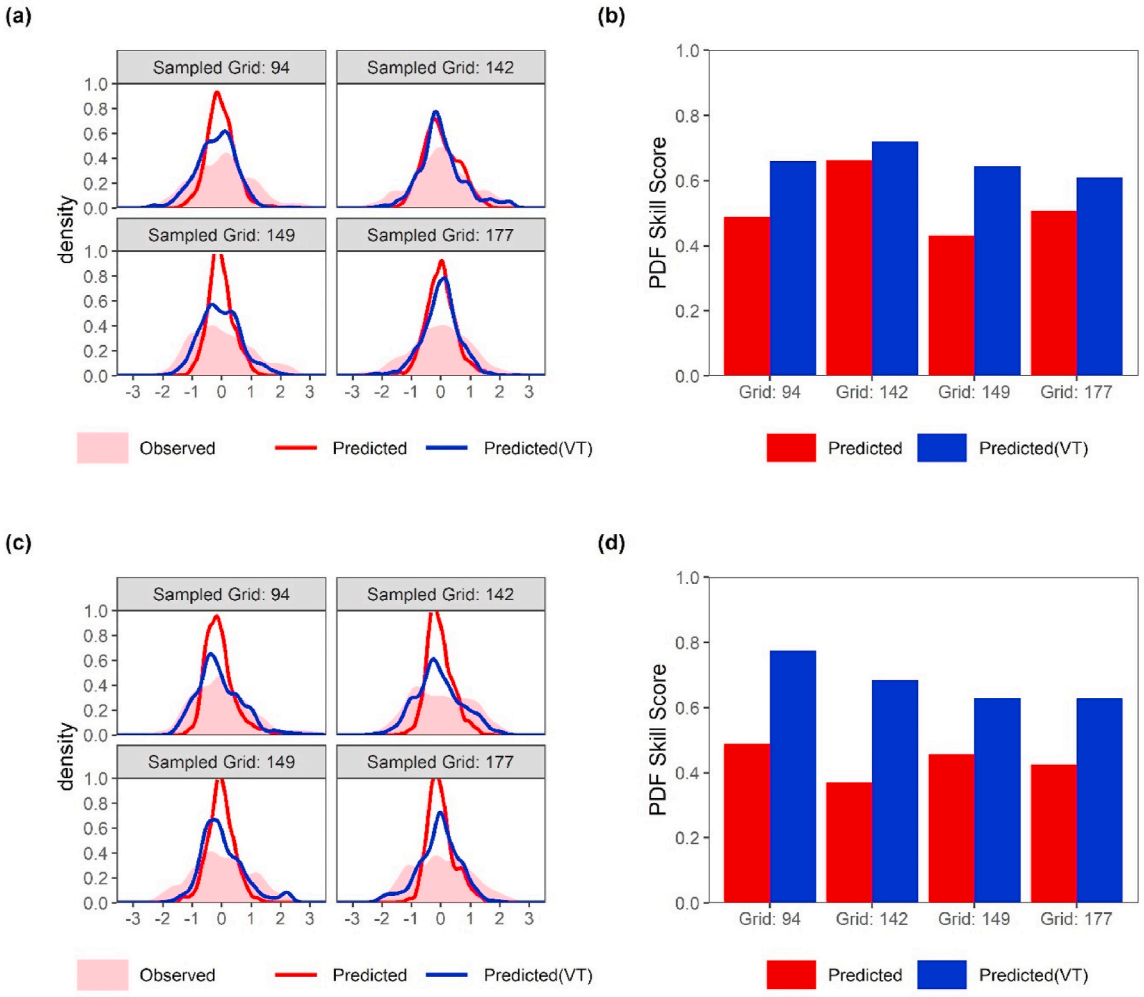

Fig. 4. Comparison between observed, predicted and predicted with variance transformation drought indices at four sampled grids. **SPI12**: **(a)** Density plot **(b)** PDF skill scores; **SPI36**: **(c)** Density plot **(d)** PDF skill scores.

Table 4
Rank of identified climate drivers by frequency at four sampled grids.

Grid	Drought Index	Model	Nino34	PDO	SAM	DMI
94	SPI12	Original	1	3	4	2
		VT	4	1	2	3
	SPI36	Original	1	4	2	3
		VT	1	2	3	4
142	SPI12	Original	1	–	–	2
		VT	1	3	2	4
	SPI36	Original	1	3	2	–
		VT	1	2	3	4
149	SPI12	Original	1	2	–	–
		VT	1	4	2	3
	SPI36	Original	–	1	–	2
		VT	1	2	3	4
177	SPI12	Original	2	1	3	–
		VT	1	2	3	4
	SPI36	Original	–	2	3	1
		VT	1	2	3	4

can capture sustained drought/wet anomalies well. A close look at the selection results in Table 4 provides more information about the benefits of the proposed method. First, additional climate indices can be selected, which is likely to result in considerable improvements at all grids. Second, even when the same predictor variables are selected (as the case in Grid 94 for both SPI12 and SPI36), the variance transformation leads to improved characterisation of sustained anomalies. Only a small improvement is observed at Grid 142 for SPI12 after applying variance transformation because at this location reasonably good skill was obtained from the original predictors.

In Fig. 5 (a) and (c), the improvement in PDF skill scores (percentage relative to non-wavelet models) for both SPI12 and SPI36 over Australia is presented. The wavelet-based method provides improvements at around 99% and 97% of grids for SPI12 and SPI36, respectively. Grids with white areas represent grids with missing data located in the central and western desert of Australia (Jones et al., 2009), while grids with black dots refer to locations with no improvements after variance transformation is used. Further, scatterplots in Fig. 5 (b) and (d) provide the magnitude of PDF skill scores at all grids over Australia, and the model using the proposed variance transformation technique outperforms the reference model using original climate indices. The biggest improvements tending to occur for locations that had lower skill with the non-wavelet model consistent with the results discussed above for Grid 142. It is noted that the improvements in prediction performance of SPI36 are larger than for SPI12, which results from possibly identifying and characterising one of the known major drivers of droughts in Australia (i.e., ENSO) using variance transformed climate indices.

What we have shown here represents the results of the MODWT-based biased variance transformation, with the results using the unbiased estimator being given in Figure S 4 of the Supporting Material. In addition, boxplots in Fig. 6 compare the model performance between approaches using biased and unbiased estimator. First, the unbiased

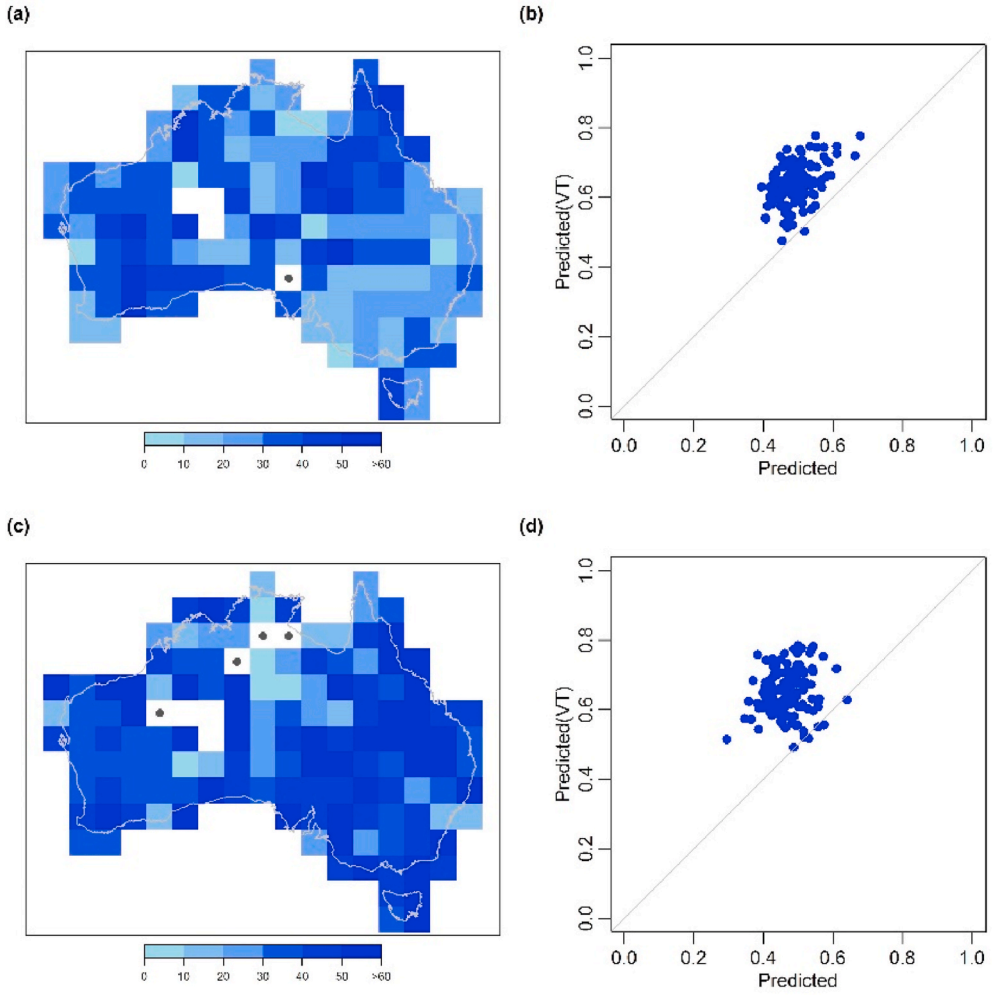

Fig. 5. Comparison of PDF skill scores between original and variance transformed (VT) predictors with MODWT-based biased variance transformation. **SPI12**: **(a)** The percent improvement of PDF skill scores over space **(b)** Scatterplot of PDF skill scores; **SPI36**: **(c)** The percent improvement of PDF skill scores over space **(d)** Scatterplot of PDF skill scores.

variance transformation does show improved prediction accuracy with all grids presenting improvements while with the biased variance transformation there are several grids perform worse than the reference model. Second, the unbiased variance transformation shows better mean statistics in both drought indices with greater improvements in SPI36. There is no significant difference in the two, which is due to the fact that the Haar wavelet filter has been used and large sample size is available in this case study.

Meanwhile, we have also done the experiments under two folds cross-validation with varying wavelet filter length seen in Table 5. The results of PDF skill scores show that first, the unbiased variance transformation approach outperform its alternative in both mean and median statistics; second, larger differences between two estimators are observed when we adopt wider wavelet filters in both mean and median statistics given the similar standard deviation (SD) across all grids. It should be noted that there is an exception when using d8 for SPI12 (median) and SPI36 (mean) prediction the difference of statistic gets smaller, which is likely due to the violation of additive decomposition when other wavelet filters are adopted. However, the results we show here confirm the argument that MODWT or AT can be applied as the basis for variance transformation even when wavelet filters other than the Haar are adopted.

4. Summary and conclusions

The open-source WASP R-package contains the codes, sample datasets and help-files for natural system prediction. We introduce the MODWT-based variance transformation, which resolves the issues of future dependence. Moreover, the boundary related bias is addressed using a newly proposed unbiased variance transformation. Both improvements have broadened the application of wavelet-based variance transformation method. The use of the wavelet-based variance transformation technique is demonstrated by predicting a drought index over Australia using various climate indices, but the logic represents a generic approach not limited to modelling hydro-climatological systems alone. This approach has shown substantial improvements in predictive accuracy especially in systems where the response and plausible predictor variables have large differences in their spectrums.

It is worth noting that this method provides a way to predict a target response in a complex system without making assumptions and simplifications including characterising the form of the underlying model that relates the two. This is implicitly undertaken by the variance transformation technique thereby formulating a transformed predictor that can be expected to have a concurrent relationship with the response ensuring improvement of predictivity in the complex system. However, the proposed approach has its inherent limitations and should be applied with care. First, the boundary related bias is a curse, thus the selection of wavelet family and the length of filters should be realistic given the

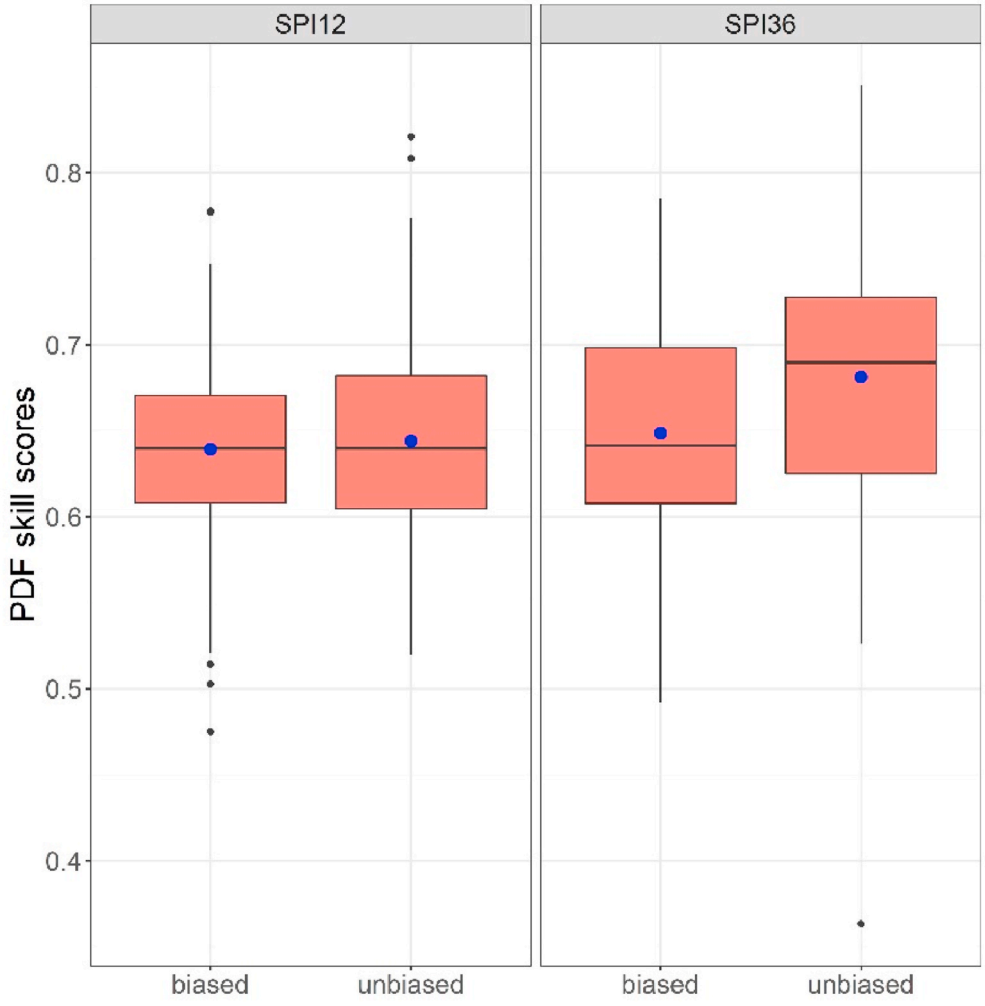

Fig. 6. Comparison of model performance between approaches using biased and unbiased estimator. Blue dots represent the mean value of PDF skill scores. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Table 5
Comparison of biased and unbiased variance transformation approach with varying wavelet filter length.

Metric	SPI	Wavelet filter	Decomposition levels (*J*)	Biased	Unbiased	Difference (Unbiased-Biased)
Mean	12	Haar(d2)	9	0.664	0.672	0.008
		d8	8	0.658	0.685	0.027
		d16	7	0.635	0.652	0.017
	36	Haar(d2)	9	0.679	0.699	0.020
		d8	8	0.716	0.718	0.002
		d16	7	0.697	0.719	0.022
Median	12	Haar(d2)	9	0.659	0.678	0.019
		d8	8	0.664	0.680	0.016
		d16	7	0.632	0.653	0.021
	36	Haar(d2)	9	0.684	0.703	0.019
		d8	8	0.703	0.724	0.021
		d16	7	0.698	0.726	0.028
SD	12	Haar(d2)	9	0.065	0.068	0.003
		d8	8	0.060	0.075	0.015
		d16	7	0.065	0.071	0.006
	36	Haar(d2)	9	0.069	0.072	0.003
		d8	8	0.064	0.068	0.004
		d16	7	0.068	0.070	0.002

nature of the physical phenomenon with varying data length (Bakshi, 1999; Maheswaran and Khosa, 2012; Percival and Walden, 2000; Torrence and Compo, 1998). In addition, the rule of thumb of the decomposition level by Kaiser (2010) is preferred such that the variance transformation is done across the entire spectrum of the predictor variables (Jiang et al., 2020).

Lastly, while the logic presented here focusses on the modelling of a single response, extensions to modelling multiple responses are possible. Future extensions of the proposed logic will illustrate how we can extend the approach here to multiple response variables, while keeping the

dimensionality of the predictive system small enough to maintain robustness in predictions.

Software availability

The open-source R-package WASP is available for download from the following website http://www.hydrology.unsw.edu.au/software/WASP and results in this work are fully reproducible through the Rmarkdown in the vignettes of this R package. Source codes are available, along with help-files and example datasets used to generate the outcomes reported.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was funded by the Australian Research Council linkage grant (LP150100548) and Crown lands & Water Division, Department of Industry, NSW, Australia. Monthly Nino3.4, PDO, and DMI are derived from monthly SST values of Hadley Centre Global Ice and Seal Surface Temperature (HadISST) dataset; SAM is calculated using SLP from NOAA Earth System Research Laboratory's Physical Sciences Division (PSD). Gridded rainfall data is obtained from the Australian Water Availability Project (AWAP) led by the Bureau of Meteorology. Assistance from Raj Mehrotra in preparing the datasets used over Australia and instructions from Jingwan Li on building R library are gratefully acknowledged. We are also thankful to two anonymous referees, Professor Holger Maier and the editor for their constructive comments.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envsoft.2020.104907.

References

Bakshi, B.R., 1999. Multiscale analysis and modeling using wavelets. J. Chemometr. 13 (3–4), 415–434.
Cai, W., Cowan, T., 2013. Southeast Australia autumn rainfall reduction: a climate-change-induced poleward shift of ocean–atmosphere circulation. J. Clim. 26 (1), 189–205.
Cai, W., Van Rensch, P., Cowan, T., Hendon, H.H., 2011. Teleconnection pathways of ENSO and the IOD and the mechanisms for impacts on Australian rainfall. J. Clim. 24 (15), 3910–3923.
Cornish, C.R., Bretherton, C.S., Percival, D.B., 2006. Maximal overlap wavelet statistical analysis with application to atmospheric turbulence. Boundary-Layer Meteorol. 119 (2), 339–374.
Dai, A., 2013. Increasing drought under global warming in observations and models. Nat. Clim. Change 3 (1), 52–58. http://www.nature.com/nclimate/journal/v3/n1/abs/nclimate1633.html#supplementary-information.
Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theor. 36 (5), 961–1005.
Du, K., Zhao, Y., Lei, J., 2017. The incorrect usage of singular spectral analysis and discrete wavelet transform in hybrid models to predict hydrological time series. J. Hydrol. 552, 44–51.
Fahimi, F., Yaseen, Z.M., El-shafie, A., 2017. Application of soft computing based hybrid models in hydrological variables modeling: a comprehensive review. Theor. Appl. Climatol. 128 (3–4), 875–903. https://doi.org/10.1007/s00704-016-1735-8.
Fierro, A.O., Leslie, L.M., 2013. Links between central west Western Australian rainfall variability and large-scale climate drivers. J. Clim. 26 (7), 2222–2246.
Fowler, H.J., Blenkinsop, S., Tebaldi, C., 2007. Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling. Int. J. Climatol. 27 (12), 1547–1578. https://doi.org/10.1002/joc.1556.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning, vol. 1. Springer series in statistics, New York, NY, USA.
Galelli, S., Humphrey, G.B., Maier, H.R., Castelletti, A., Dandy, G.C., Gibbs, M.S., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Software 62, 33–51. https://doi.org/10.1016/j.envsoft.2014.08.015.
Hertig, E., Tramblay, Y., 2017. Regional downscaling of Mediterranean droughts under past and future climatic conditions. Global Planet. Change 151, 36–48.
Jiang, Z., Sharma, A., Johnson, F., 2019. Assessing the sensitivity of hydro-climatological change detection methods to model uncertainty and bias. Adv. Water Resour. 134, 103430 https://doi.org/10.1016/j.advwatres.2019.103430.
Jiang, Z., Sharma, A., Johnson, F., 2020. Refining predictor spectral representation using wavelet theory for improved natural system modeling. Water Resour. Res. 56 (3), e2019WR026962 https://doi.org/10.1029/2019WR026962.
Jones, D.A., Wang, W., Fawcett, R., 2009. High-quality spatial climate data-sets for Australia. Aust. Meteorol. Oceanogr. J. 58 (4), 233.
Kaiser, G., 2010. A Friendly Guide to Wavelets. Springer Science & Business Media.
Kirono, D.G., Chiew, F.H., Kent, D.M., 2010. Identification of best predictors for forecasting seasonal rainfall and runoff in Australia. Hydrol. Process.: Int. J. 24 (10), 1237–1247.
Lall, U., Sharma, A., 1996. A nearest neighbor bootstrap for resampling hydrologic time series. Water Resour. Res. 32 (3), 679–693.
Maheswaran, R., Khosa, R., 2012. Comparative study of different wavelets for hydrologic forecasting. Comput. Geosci. 46, 284–295. https://doi.org/10.1016/j.cageo.2011.12.015.
McKee, T.B., Doesken, N.J., Kleist, J., 1993. The relationship of drought frequency and duration to time scales. In: Paper Presented at the Proceedings of the 8th Conference on Applied Climatology.
Mehrotra, R., Sharma, A., 2006a. Conditional resampling of hydrologic time series using multiple predictor variables: a K-nearest neighbour approach. Adv. Water Resour. 29 (7), 987–999.
Mehrotra, R., Sharma, A., 2006b. A nonparametric stochastic downscaling framework for daily rainfall at multiple locations. J. Geophys. Res.: Atmosphere 111 (D15).
Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012.
Murphy, B.F., Timbal, B., 2008. A review of recent climate variability and climate change in southeastern Australia. Int. J. Climatol. 28 (7), 859–879.
Nason, G.P., 2008. Wavelet Methods in Statistics with R. Springer, New York: New York. New York.
Ndehedehe, C.E., Agutu, N.O., Okwuashi, O., Ferreira, V.G., 2016. Spatio-temporal variability of droughts and terrestrial water storage over Lake Chad Basin using independent component analysis. J. Hydrol. 540, 106–128.
Nguyen, H., Mehrotra, R., Sharma, A., 2019. Correcting systematic biases across multiple atmospheric variables in the frequency domain. Clim. Dynam. 52 (1–2), 1283–1298.
Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, Cambridge.
Perkins, S., Pitman, A., Holbrook, N., McAneney, J., 2007. Evaluation of the AR4 climate models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. J. Clim. 20 (17), 4356–4376.
Pui, A., Sharma, A., Santoso, A., Westra, S., 2012. Impact of the El Niño southern oscillation, Indian Ocean Dipole, and southern annular mode on daily to sub-daily rainfall characteristics in east Australia. Mon. Weather Rev. 140, 1665–1681. https://doi.org/10.1175/MWR-D-11-00238.1.
Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. J. Hydrol. 563, 336–353.
Quilty, J., Adamowski, J., Boucher, M.A., 2019. A stochastic data-driven ensemble forecasting framework for water resources: a case study using ensemble members derived from a database of deterministic wavelet-based models. Water Resour. Res. 55 (1), 175–202. https://doi.org/10.1029/2018wr023205.
Rashid, M.M., Beecham, S., 2019. Simulation of streamflow with statistically downscaled daily rainfall using a hybrid of wavelet and GAMLSS models. Hydrol. Sci. J. 64 (11), 1327–1339.
Rashid, M.M., Beecham, S., Chowdhury, R.K., 2016. Statistical downscaling of rainfall: a non-stationary and multi-resolution approach. Theor. Appl. Climatol. 124 (3–4), 919–933.
Rashid, M.M., Johnson, F., Sharma, A., 2018. Identifying sustained drought anomalies in hydrological records: a wavelet approach. J. Geophys. Res.: Atmosphere 123 (14), 7416–7432.
Rashid, M.M., Sharma, A., Johnson, F., 2020. Multi-model drought predictions using temporally aggregated climate indicators. J. Hydrol. 581, 124419.
Rayner, N., Parker, D.E., Horton, E., Folland, C.K., Alexander, L.V., Rowell, D., Kaplan, A., 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J. Geophys. Res.: Atmosphere 108 (D14).
Sang, Y.F., 2013. A review on the applications of wavelet transform in hydrology time series analysis. Atmos. Res. 122, 8–15. https://doi.org/10.1016/j.atmosres.2012.11.003.
Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 — a strategy for system predictor identification. J. Hydrol. 239 (1), 232–239. https://doi.org/10.1016/S0022-1694(00)00346-2.
Sharma, A., Mehrotra, R., 2014. An information theoretic alternative to model a natural system using observational information alone. Water Resour. Res. 50 (1), 650–660.
Sharma, A., Mehrotra, R., Li, J., Jha, S., 2016. A programming tool for nonparametric system prediction using Partial Informational Correlation and Partial Weights. Environ. Model. Software 83, 271–275.
Sharma, A., Tarboton, D.G., Lall, U., 1997. Streamflow simulation: a nonparametric approach. Water Resour. Res. 33 (2), 291–308.
Sheffield, J., 2011. Drought : Past Problems and Future Scenarios. Earthscan, London Washington, DC: London Washington, DC.
Sheffield, J., Wood, E.F., 2008. Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, IPCC AR4 simulations. Clim. Dynam. 31 (1), 79–105.

Spinoni, J., Naumann, G., Carrao, H., Barbosa, P., Vogt, J., 2014. World drought frequency, duration, and severity for 1951–2010. Int. J. Climatol. 34 (8), 2792–2804.

Strang, G., 1996. Wavelets and Filter Banks. Wellesley-Cambridge Press, Wellesley, MA: Wellesley, MA.

Torrence, C., Compo, G.P., 1998. A practical guide to wavelet analysis. Bull. Am. Meteorol. Soc. 79 (1), 61–78. https://doi.org/10.1175/1520-0477(1998)079<0061:Apgtwa>2.0.Co;2.

Weinberger, K.Q., Blitzer, J., Saul, L.K., 2006. Distance metric learning for large margin nearest neighbor classification. In: Paper Presented at the Advances in Neural Information Processing Systems.

Westra, S., Sharma, A., 2010. An upper limit to seasonal rainfall predictability? J. Clim. 23 (12), 3332–3351. https://doi.org/10.1175/2010JCLI3212.1.