Forecasting the daily variability discharge in the fluvial system of the Paraná River. An ODPC hydrology application

Melanie Meis[*], Manuel Benjamin, Daniela Rodriguez
Universidad de Buenos Aires and CONICET.

Abstract

As it was stated in the last report from the Intergovernmental Panel on Climate Change (IPCC) extreme hydrological events have increased in the last decades and models forecast forsee that those events will get worse in the near future. Floods and downspouts are typical extreme events that develop in the fluvial system of the La Plata Basin, particularly in the Paraná's river. The development of those events affect the socio-economic areas in the countries that surround the river. In this sense, studies on new resources to prevent and reduce those impacts are needed. We proposed to apply a new method in the hydrology field known as One Sided Dynamical Principal Components (ODPC) to model and predict the daily variability from one and three days of the streamflow of the Paraná River. We did a comparison between ODPC and the common models used in the research field through different indices as the root mean squared error; Nash-Sutcliffe index, among others. This comparison with standard time series models demonstrates the consistency of the ODPC method for the treatment and forecasting of the daily variability discharge over the other models especially for the three-day forecast.

Keywords. ODPC - Streamflow -Modeling- Forecasting - Multivariate

[*] Corresponding author: meis.melanie@cima.fcen.uba.ar

1 Introduction

Extreme events such as floods and droughts might increase as it was stated in the last IPCC report. New tools to prevent and manage those events are needed (Vanelli and Kobiyama (2021)). In this sense, a good use of the hydrological resources from a country implies to provide in consequence an adequate management of the river transport, hydroelectric energy, tourism, fishing and water supply in the different scenarios that it might possibly occur. In this context, it is necessary to have a better knowledge of the variability that occurs in the discharges in order to contribute with valuable information to decision makers.

Since climate change is a significative degradation driver in the majority of the coastal systems affecting its diversity, a short and long term planification of the social and economic activities needs to be done (Bridgewater and Schmeller (2018); Talebmorad et al. (2021); Javadinejad et al. (2021)). Moreover, various types of extreme hydrology-climatology events are predicted to occur more frequently with potentially strong implications in coastal areas in many regions around the globe (FAO and Hunger (2015)). Particularly, rapid floods are expected to rise in frequency and severity due to the impact of global change in the climate system (Huntington (2006),; Borga et al. (2011); Ostad-Ali-Askar et al. (2018)).

The development and application of new statistical techniques that might improved the runoff forecasting have been of constant interest for the scientific and non-scientific community in the last decades due to the impact of Climate Change and its implication in the water supply (Zubaidi et al. (2020); Vanelli and Kobiyama (2021). Different works have shown hydro-informatic techniques with computationally low cost could be applied to deal with hydrology negative aspects, especially in developing countries (Le et al. (2019)). Besides, in vulnerable and densely populated countries, Kundzewicz (2002) noticed that floods continue being the principal cause of great deaths in relation to other natural disasters.

It is well known that hydrology is a data-intensive field, in which massive data accumulated by years contain a lot of information, and advanced technology nowadays (Astagneau et al. (2021)) enables for the collection of this type of multivariate data. On the other hand, a large volume of literature exists on hydrological time series, most of which focuses on time series prediction, Porporato and Ridolfi (2001); Niu and Feng (2021); Fathian (2021), among others.

In particular, multivariate time series analysis has many applications, as it can account for interrelations between several variables (Wei (2018)) . However, these models face the challenge of complexity in their structures, even when modeling series with large dimensions. This complexity occurs because the number of the parameters expands enormously fast as the dimension increases. Therefore, reducing the dimension of the series becomes critical to manage such data. An usual strategy to avoid the multivariate time series consists in applying a reduction of the dimension transforming the data and then applying a univariate time series model to the transformed data (see for instance Westra et al. (2007)). The Principal component analysis (PCA) is a multivariate statistical technique for data dimensionality reduction and features extraction. However, the conventional PCA can not fully reveal the entire variation and relations of the data when they are serially correlated (Alshammri and Pan (2021), Brillinger (1964)).

The Dynamic principal component analysis (DPCA) was originally proposed by Brillinger (1981). Furthermore, the main advantage of the DPCA is that they generalize the classical PCA by considering serial dependence of data and they reduce (referring to the components) the dimension of time series in the frequency domain while preserving as much information as possible. Peña et al. (2019), proposed an empirical approach to named one-sided dynamic principal components (ODPC) for time series. This approach is defined as linear combinations of the present and past values of the series that minimize the reconstruction mean squared error. Previous definitions of DPCA depended on past and future values of the series. For this reason, they are not appropriate for forecasting purposes.

Streamflow forecasting could be established in two categories: short-term and long-term forecast. The first one regards to hourly and daily streamflow predictions, while the second one to weekly, monthly or annual ones, Yaseen et al. (2015). In this present work a short-term forecast was considered. Particularly, the aim of this research was to apply a statistical method, not used before in the hydrology field, to model and forecast the daily variability's discharge and see if an improvement could be achieve in front of the classical statistical methodologies apply. Thus, to assert the robustness of the new model, we compared the ODPC method with standard time series procedures through a wide range of indices like the root mean squared error (RMSE), the Nash-Sutcliffe index and the mean absolute error (MAE).

2 Data

In this work, we considered the daily mean discharge data of the Integrated Hydrology data base (BDHI) of the Under Subsecreteriat of Hydrologycal Resources of the Argentine Repúblic. We analysed the data from Corrientes and Itatí (Paraná river), Puerto Pilcomayo (Paraguay river) covering the time period from 26/11/1999 to 31/12/2015. In Figure 1 the location of the gauging stations is shown.

All the time series analyzed presented less than 10% of missing data. However, in order to complete their nan-values we considered linear interpolation only if there were at least two consecutive days. Otherwise, the most similar month of another year was contemplated to complete the missing points. The most similar month also contemplated the similarity to the four months before and after the missing data points.

Furthermore, as it is well known the streamflow's time series present a large temporal variability. In this sense, in order to stabilize the variance, we decided to evaluate the logarithm in the data (Figure 2) and then we differentiated the daily discharge into one and three orders depending on whether the forecast frontier was one or three days ahead. This means that we estimated the difference between the flow of a given day with respect to the day of the forecast horizon.

Particularly, we were interested in forecasting Corrientes's variability discharge. The location of this gauge station is essential for economic and social purposes in the region. Even more, Pilcomayo and Itatí gauge stations are essential to describe the variability of the discharge measured in Corriente's gauge station as both are upstream it, Battistello Espíndola and Ribeiro (2020).

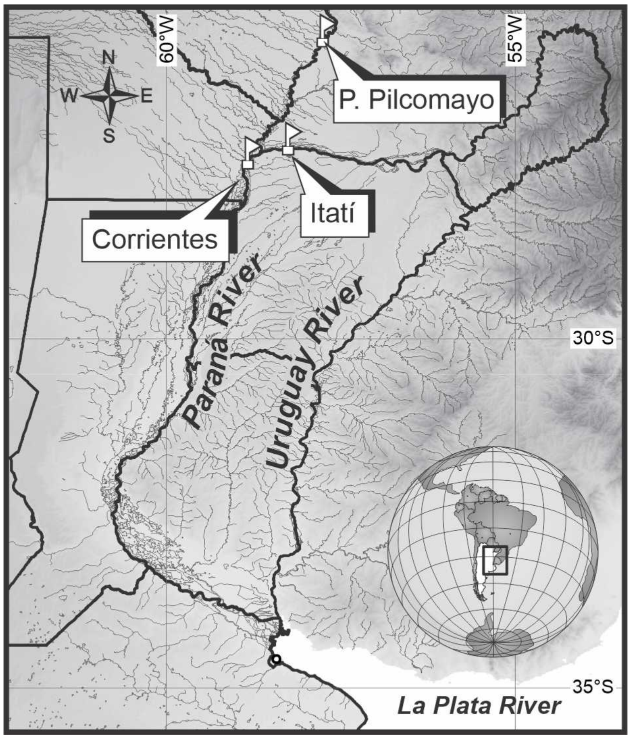

Figure 1: Location from the gauge discharge stations in the La Plata Basin.

3 Models

In this section we summarized the models that were considered for modeling and forecasting the mean daily variability flow. First we introduced the One Sided Dynamic Principal Components (ODPC) and discussed their advantages for forecasting large sets of time series. Then we considered three time series models that correspond to different models studied in the literature in the context of forecast mean daily flow and compared their performance against ODPC.

As we mentioned in the introduction, The Dynamic principal component analysis (DPCA) has been developed by Brillinger (1981) to reduce dimension in multivariate time-series data. The main advantage of DPCA is its capability of extracting essential components from the data by reflecting the serial dependence on them. In this work we applied the methodology one sided dynamic principal components (ODPC) Peña et al. (2019), which computes a linear combination of present and previous values of the time series that minimizes the mean squared error of reconstruction. These components are a modification of Brillinger (1964,

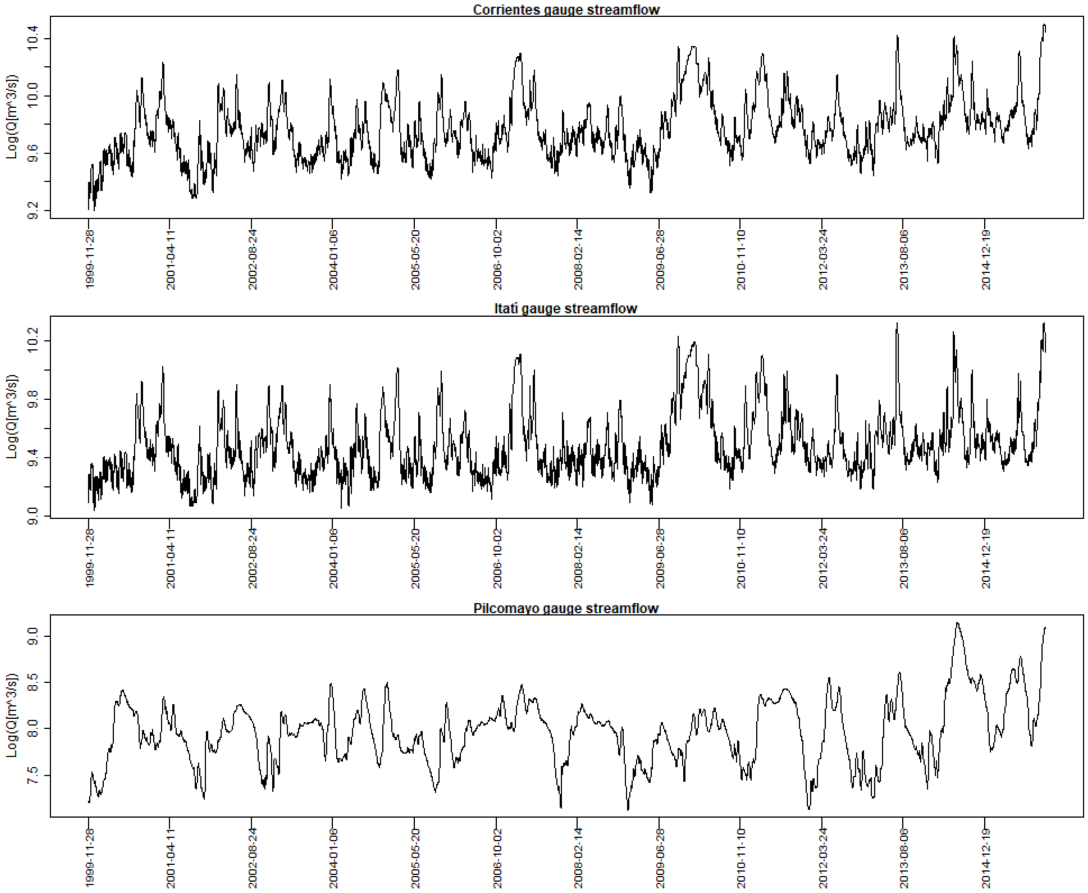

Figure 2: Discharge time series (m^3/s) a) Corrientes, b) Itatí y c) Pilcomayo.

1981) dynamic principal components. The main difference is that the latter uses also future values to reconstruct the time series. This makes ODPC more adequate for forecasting purposes and enables the reduction of the dimension of multivariate time series. The ODPC conserves the spirit of principal component analysis, without assuming any particular model (parametric or not) for the data. Also the fact that it is not based on both lags and leads of the data, makes it useful for forecasting large sets of time series.

The ODPC methodology has the principal aim to find a vector for the time series with less dimension that manages to represent the dynamical structure of the data.

More precisely, denoted with $Z_t \in R^m$ the time series observed and considered $a = (a_0^{'}, ..., a_k^{'}) \in R^{m\times(k+1)}$ with $\|a\| = 1$, $B \in R^{(k+1)\times m}$ and $(B)_{hj} = b_{h,j}$. We defined the mean squared error in the reconstruction of the data by the following equation

$$MSE(a,B) = \frac{1}{T-(k_1+k_2)} \sum_{t=k_1+k_2+1}^{T} \sum_{j=1}^{m} (Z_{t,j} - \sum_{h=0}^{k_2} b_{h,j} f_{t-h}(a))^2 \tag{3.1}$$

where

$$f_t(a) = \sum_{h=0}^{k_1} a_h^{'} Z_{t-h}, \quad t = k+1, ..., T \tag{3.2}$$

and k_1 and k_2, tunning parameters to be chosen. Hence, we defined the estimators $\hat{a}$ and $\hat{B}$ that minimize the $MSE(a,B)$. By this form the principal dynamic components are obtained as:

$$\hat{Z}_t = \sum_{h=0}^{k_2} \hat{b}_h f_{t-h}(\hat{a}) \tag{3.3}$$

We can observe that the dynamic principal components are defined as a linear combination of the past and present values of the time series that minimizes the reconstruction of the mean squared error. The number of lags (k_i, with $i = 1, 2$) considered in the reconstruction is parameter to be estimated and should be chosen. The k_1 is the number of lags used to define the principal dynamic component, while k_2 is the number of lags used for the reconstruction. We considered the cross-validation implemented by Peña et al. (2019), for the selection of those parameters.

In order to compare the ODPC methodology in the context of forecasting the mean daily flow, we considered three competitors in the study: SARIMA, SARIMAX and VAR. Now we introduce the classical models used in the comparison with ODPC and we refer the reader to Box et al. (2015); Durbin and Koopman (2012); Hamilton (2020) for a fuller thorough treatment of the models.

SARIMA: The SARIMA models can study time series that do not present a stationary process, and can include seasonality. These models are represented in this form:
SARIMA$(p,d,q)(P,D,Q)_S$, where (p,d,q) is the ordinary part, while $(P,D,Q)_S$ is the seasonal part (implemented by Hyndman, 2016). In general, the model can be expressed as follows (Eq. 2.1):

$$\phi_{PS}(B_S)\phi_p(B)(1-B_S)^D(1-B)^d Z_t = \theta_{QS}(B_S)\theta_q(B)u_t \tag{3.4}$$

SARIMAX: Including external variables as input to improve the performance of a SARIMA model is known as SARIMAX (Xie et al., 2013). The model can be represented by the following equation (Eq. 2.2):

$$\phi_P(B_s)\phi_p(B)(1-B_s)^D(1-B)^d Z_t = \theta_Q(B_s)\theta_q(B)u_t + \sum_{h=0}^{b}\beta_h X_{t-h} \tag{3.5}$$

With $D > 0$ is the difference order associated to the seasonal part of the model, and X_t is the external variable.

VAR: The autoregressive model vector (VAR) is an extension of the univariate autoregressive model. This kind of model is useful when we are interested in a forecast to predict multiple time series considering only one model. In particular, for this analysis we followed the procedure established by Praff (2008), where the model was estimated considering regression linear methods for each equation. The statistical method can be studied by the equation 2.3:

$$Z_t = A_1 Z_{t-1} + \ldots + A_p Z_{t-p} + CX_t + u_t \tag{3.6}$$

The Z_t is an endogenous vector variable of $K \times 1$, and u_t represents a spherical perturbation of same dimension. The matrix $A_1, .., A_p$ is dimension $K \times K$. Besides, a constant or a trend could be included in the deterministic model, as well as a seasonal dummy or exogenous variable (CX_t).

We use cross-validation procedures for each model to select their hyper parameters. The last two thousand observations are used as tests points in order to evaluate each model prediction performance. In this sense, all the models were fitted with observations that occurred previously to the observation considered for the testing value. In Figure 3 it is schematized the evaluation of the forecasting procedure. The blue points correspond to the training observations while the red points are the test observations Hyndman and Koehler (2006).Then, we estimated the RMSE (Eq. 3.7) , NSE (Eq. 3.8) and MAE (Eq. 9) metrics along the two thousand residuals obtanied for each model.

The whole modeling and forecasting procedure could be resume in the following flow chart (Figure 4).

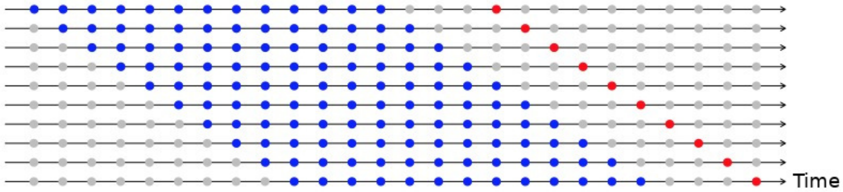

Figure 3: Forecasting testing procedure. Blue points correspond to training observations while the red points are the test observations.

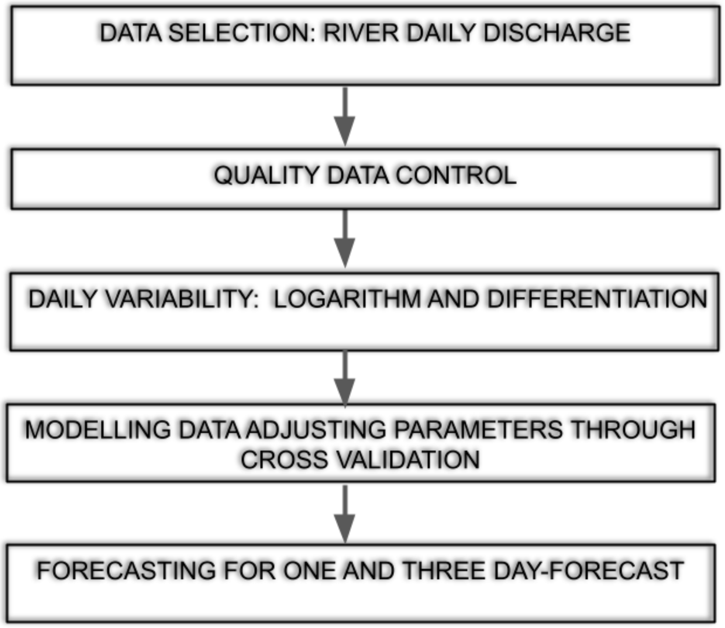
DATA SELECTION: RIVER DAILY DISCHARGE
QUALITY DATA CONTROL
DAILY VARIABILITY: LOGARITHM AND DIFFERENTIATION
MODELLING DATA ADJUSTING PARAMETERS THROUGH CROSS VALIDATION
FORECASTING FOR ONE AND THREE DAY-FORECAST

Figure 4: Flow chart for the modeling and forecasting procedure.

$$\text{RMSE} = \sqrt{\frac{1}{2000}\sum_{t=1}^{2000}(Y_t - \widehat{Y}_t)^2} \tag{3.7}$$

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{2000}(\widehat{Y}_t - Y_t)^2}{\sum_{t=1}^{2000}(Y_t - \bar{Y}_t)^2} \tag{3.8}$$

$$\text{MAE} = \frac{1}{2000}\sum_{t=1}^{2000}\left|\widehat{Y}_t - Y_t\right| \tag{3.9}$$

Where $\bar{Y}_t$ is the mean observed discharge, $\widehat{Y}_t$ is the modeled discharge, and Y_t is the observed discharge at time t.

Finally, we estimated the mean squared error (MSE) curves for each method. Let Y_t

with $1 \leq t \leq N$ be the corresponding values to be estimated then their predictions $\widehat{Y}_t$ are obtained with all previous data points. We defined for each $1 \leq k \leq 2000$

$$\mathrm{MSE}(k) = \frac{1}{k} \sum_{t=1}^{k} (Y_t - \widehat{Y}_t)^2 \tag{3.10}$$

4 Results

We were interested in studying and forecasting the daily discharge variability for Corrientes gauge station. We started by taking the logarithm of the time series. For the one-day forecast, we differentiated the time series once with *d=1*, while for the three-day forecast the differentiation was done with *d=3*. Then, we ran all the four models and we estimated their mean squared prediction error. We observed the presence of two estimated errors that were far away from the mean dispersion of the whole errors, and in particular they were introducing certain noise to all the forecast estimations. Both errors were the consequence of two atypical points of the time series differences, we decided to replace them by a linear interpolation. This procedure was done in the three-time differentiation time series too. After replacing the atypical points we ran the models again with the procedure already mentioned in the methodology section.

We tested each methodology with the last $N = 2000$ data points of the multivariate time series. In Table I and II we present Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and Nash-Sutcliffe model Efficency Coefficient (NSE) for each model.

In the one-day forecast, the SARIMA model exhibited the worst performance which was far separated from the rest of the models. This, could be attributed to its inability to make use of the information from other gauging stations.

It can be seen that the ODPC model presented a better performance under all metrics. Particularly, the best improvement was regarding the univariate models. Furthermore, the benefits of the ODPC model are better appreciated in the three-day forecast as the ODPC model shows the best relative performance, and where the NSE exhibited the highest value.

The MSE curves (Eq. 3.10) for each methodology are presented in Figure 5 and Figure 6. A certain tendency can be observed in all the curves so it may not be correct to claim

	ODPC	VAR	SARIMA	SARIMAX
RMSE	0.011	0.012	0.015	0.012
MAE	0.008	0.008	0.011	0.009
NSE	0.726	0.717	0.537	0.692

Table 1: Corrientes gauge station, Parana River one day forecast horizon.

	ODPC	VAR	SARIMA	SARIMAX
RMSE	0.041	0.045	0.050	0.051
MAE	0.030	0.034	0.039	0.039
NSE	0.460	0.341	0.185	0.154

Table 2: Corrientes gauge station, Parana River three day forecast horizon.

that each methodologie's expected squared error has been correctly estimated in the last prediction. Still it is arguable that the ODPC methodology is better in terms of prediction error since it presents a uniformly lower MSE curve than the rest of the methodologies in both one and three day forecasts.

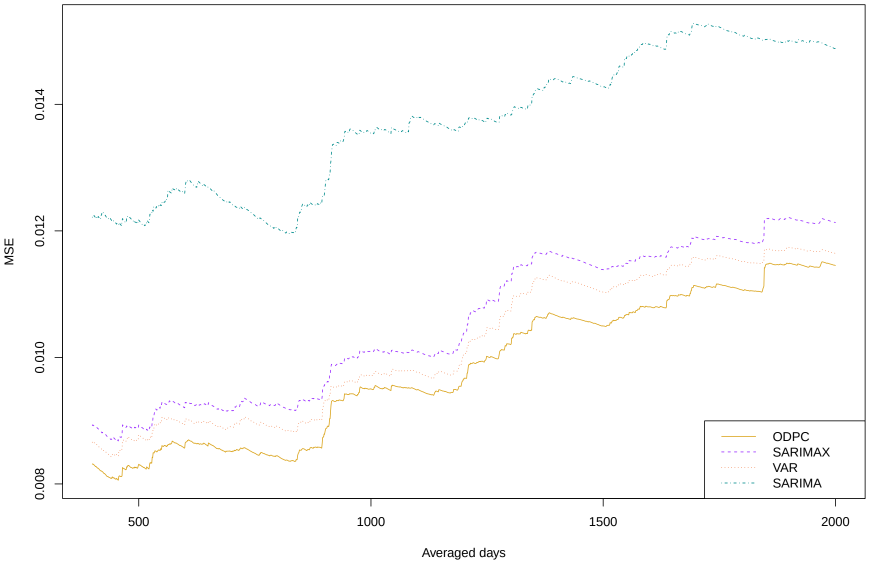

Figure 5: Mean squared error for one-day forecast

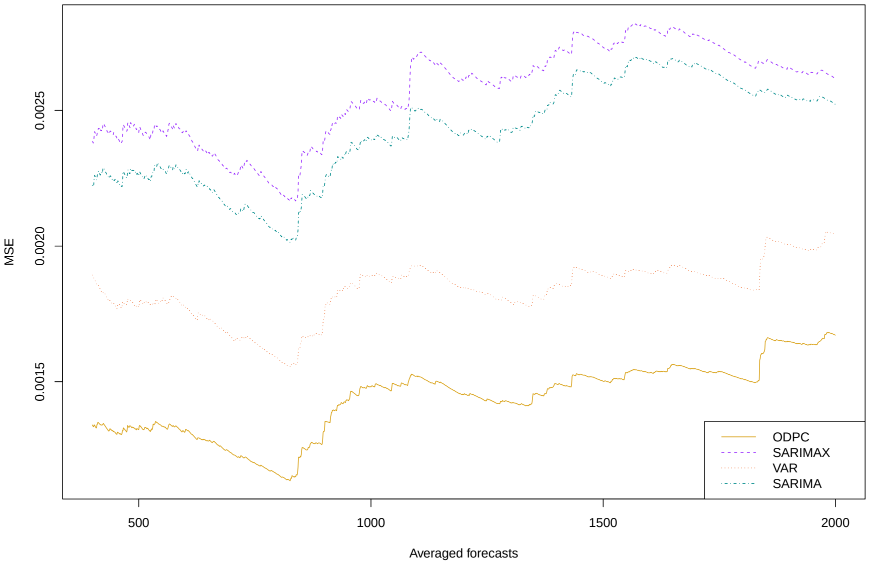

Figure 6: Mean squared error for three-day forecast

5 Conclusion

Climate change is significative affecting human social and economic life. Particularly, coastal areas and its diversity are being influenced by floods and down-spouts. In this sense a short and long term planification of the social and economic activities needs to be done Biao (2017). Different approaches to model and forecast discharge are necessary due to the stochastic behaviour from the variable itself.

A wide range of statistical models can represent more or less efficient the behavior of different time series, but most of them lack an understandable explanation for their development. In addition, the understanding of the models' results by decision makers is essential to decide the appropriate decisions for the management of natural resources.

Particularly, in this work we decided to explore several time series models to evaluate the prediction of the time series variability of the River Basin Of the Paraná River. The region considered is important since a large part of the economy of the surrounding countries depends on it.

In this sense, we considered a forecasting horizon of one and three days ahead. We evaluated the performance of a new statistical method applied in the hydrology field, the ODPC method. The methodology exhibited the best performance metric comparison to the all models applied. Even more, through cross validation we found that the ODPC methodology was essential to reduce the mean squared error in one and three day forecast. Furthermore, the three-day forecast using ODPC showed greater improvements compared to the rest of the methods than the one-day forecast.

Declarations.

Funding

This research was supported by CONICET under Grant 11220130100806, 20020170100330BA from the Universidad de Buenos Aires and PICT-201-0377 from ANPYCT, Argentina.

Conflicts of interest

Authors declare no conflicts of interest.

Availability of data and material

Data used in this research are available through the web page at the Subsecretaría de Recursos Hídricos (Argentine Undersecretariat for Water Resources).

Code availability

The code used for the estimation of the discharge could be available under request. Libraries from R were applied.

Ethics approval

Not applicable.

Consent to participate

Not applicable.

Consent for publication

Not applicable.

References

Alshammri, F. and Pan, J. (2021). Moving dynamic principal component analysis for non-stationary multivariate time series. *Computational Statistics*, 36.

Astagneau, P. C., Thirel, G., Delaigue, O., Guillaume, J. H., Parajka, J., Brauer, C. C., Viglione, A., Buytaert, W., and Beven, K. J. (2021). Hydrology modelling r packages–a unified analysis of models and practicalities from a user perspective. *Hydrology and Earth System Sciences*, 25(7):3937–3973.

Battistello Espíndola, I. and Ribeiro, W. C. (2020). Transboundary waters, conflicts and international cooperation-examples of the la plata basin. *Water International*, 45(4):329–346.

Biao, E. I. (2017). Assessing the impacts of climate change on river discharge dynamics in oueme river basin (benin, west africa). *Hydrology*, 4(4):47.

Borga, M., Anagnostou, E., Blöschl, G., and Creutin, J.-D. (2011). Flash flood forecasting, warning and risk management: the hydrate project. *Environmental Science & Policy*, 14(7):834–844.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons.

Bridgewater, P. and Schmeller, D. S. (2018). Ipbes-6: the best plenary yet?

Brillinger, D. (1964). The generalization of the techniques of factor analysis, canonical correlation and principal components to stationary time series. In *Invited Paper at the Royal Statistical Society Conference in Cardiff, Wales*.

Brillinger, D. (1981). Time series: Data analysis and theory (expanded edn) holden-day. *San Francisco*.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods*. Oxford university press.

FAO, I. and Hunger, W. A. Z. (2015). The critical role of investment in social protection and agriculture. *Rome, FAO*.

Fathian, F. (2021). *Introduction of multiple/multivariate linear and nonlinear time series models in forecasting streamflow process*. Elsevier.

Hamilton, J. D. (2020). *Time series analysis*. Princeton university press.

Huntington, T. G. (2006). Evidence for intensification of the global water cycle: Review and synthesis. *Journal of Hydrology*, 319(1-4):83–95.

Hyndman, R. J. and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4):679–688.

Javadinejad, S., Eslamian, S., and Ostad-Ali-Askari, K. (2021). The analysis of the most important climatic parameters affecting performance of crop variability in a changing climate. *International journal of hydrology science and technology*, 11(1):1–25.

Kundzewicz, Z. W. (2002). Non-structural flood protection and sustainability. *Water International*, 27(1):3–13.

Le, X.-H., Ho, H. V., Lee, G., and Jung, S. (2019). Application of long short-term memory (lstm) neural network for flood forecasting. *Water*, 11(7):1387.

Niu, W. and Feng, Z. (2021). Evaluating the performances of several artificial intelligence methods in forecasting daily streamflow time series for sustainable water resources management. *Sustainable Cities and Society*, 64:102562.

Ostad-Ali-Askar, K., Su, R., and Liu, L. (2018). Water resources and climate change. *Journal of Water and Climate Change*, 9(2):239.

Peña, D., Smucler, E., and Yohai, V. J. (2019). Forecasting multiple time series with one-sided dynamic principal components. *Journal of the American Statistical Association*.

Porporato, A. and Ridolfi, L. (2001). Multivariate nonlinear prediction of river flows. *Journal of Hydrology*, 248(1-4):109–122.

Talebmorad, H., Abedi-Koupai, J., Eslamian, S., Mousavi, S.-F., Akhavan, S., Ostad-Ali-Askari, K., and Singh, V. P. (2021). Evaluation of the impact of climate change on reference crop evapotranspiration in hamedan-bahar plain. *International Journal of Hydrology Science and Technology*, 11(3):333–347.

Vanelli, F. M. and Kobiyama, M. (2021). How can socio-hydrology contribute to natural disaster risk reduction? *Hydrological Sciences Journal*, 66(12):1758–1766.

Wei, W. W. (2018). *Multivariate time series analysis and applications*. John Wiley & Sons.

Westra, S., Brown, C., Lall, U., and Sharma, A. (2007). Modeling multivariable hydrological series: Principal component analysis or independent component analysis? *Water Resources Research*, 43(6).

Yaseen, Z. M., El-Shafie, A., Jaafar, O., Afan, H. A., and Sayl, K. N. (2015). Artificial intelligence based models for stream-flow forecasting: 2000–2015. *Journal of Hydrology*, 530:829–844.

Zubaidi, S. L., Ortega-Martorell, S., Kot, P., Alkhaddar, R. M., Abdellatif, M., Gharghan, S. K., Ahmed, M. S., and Hashim, K. (2020). A method for predicting long-term municipal water demands under climate change. *Water Resources Management*, 34(3):1265–1279.