A free and state-of-the-art probabilistic flow forecasting tool designed for Africa

J.P. Matos
École Polytechnique Fédérale de Lausanne
Laboratory of Hydraulic Constructions
EPFL-ENAC-IIC-LCH
GC A3 504, Station 18
CH-1015 Lausanne, Switzerland

A.J. Schleiss
École Polytechnique Fédérale de Lausanne
Laboratory of Hydraulic Constructions
EPFL-ENAC-IIC-LCH
GC A3 504, Station 18
CH-1015 Lausanne, Switzerland

Introduction

Flood forecasting, and more broadly flow forecasting, are an extremely relevant topic worldwide. This is certainly the case in Africa, due to significant reasons. Water is a prized resource in much of the continent, being that basic needs can be better served and substantial economic gains made using effective flow forecasting techniques. Also, populations often occupy dangerous floodplain areas due to lack of awareness of the risks, scarce information, and necessity, being particularly exposed to floods; more lives and livelihoods can be saved when flood warnings are issued timely.

On the one hand flow forecasting deals with uncertainty and is a technically challenging problem over which the scientific community has invested both a lot of time and resources. Yet, despite great advancements having been made in the past, many scientific and technical challenges associated with flow forecasting are far from being solved. On the other hand, in many African countries, the expertise to maintain flow forecasting systems is lacking, the required data is simply not available, and the infrastructure needed to prepare and transmit it is not in place.

A novel state-of-the-art probabilistic flow forecasting tool that aims to address these shortcomings is presented in this contribution. Designed for Africa and based on machine learning techniques, the tool is cheap to implement (in fact nearly free, as the code is open-sourced), has fully adaptable data requirements (simulations improve as more relevant data is used as a predictor), and provides a full and extremely accurate depiction of forecast uncertainty. Presently, the system is being tested on the Zambezi River Basin, where it is expected to alleviate the impacts of flooding and contribute to enhance hydropower production.

1. Background

1.1 Motivation

Flow forecasting is paramount for reservoir management and plays a central role in the prevention of downstream floods and the optimization of hydropower production. Accurate flow forecasting is, however, not always easy. In fact, in order to obtain flow forecasts one should have access to numerical weather forecasts, a calibrated hydrological model, and an ensemble forecasting system that brings those components together [*Cloke and Pappenberger*, 2009].

Even if all those components are available, it should be kept in mind that forecasts are always uncertain, and it is therefore important to have a good idea about the quality of the information that is being relied upon to make reservoir management decisions. One way to do so is to move from a deterministic paradigm to a probabilistic one. In fact, that has been the tendency of the hydrology community for several years.

Obtaining reliable probabilistic forecasts can be a complex and computationally intensive process. Arguably, the two main ways to achieve that goal are uncertainty postprocessors or ensemble forecasting systems. Uncertainty postprocessors rely on a deterministic hydrological model whose predictive error is modeled based on past forecasting performance and can, therefore, be estimated operationally [*Krzysztofowicz*, 2002; *Todini*, 2008; *Solomatine and Shrestha*, 2009; *Weerts et al.*, 2011]. Ensemble forecasting systems are based on multiple runs of one or more deterministic models among which sets of parameters, initial conditions, and inputs change. As such, their reliability depends, firstly, on the reliability of the numerical weather prediction system and, secondly, on the data assimilation scheme (*e.g.* an Ensemble Kalman Filter) that manages each model run. Often, the predictive

distributions produced by ensemble forecasting systems must themselves be postprocessed to match observations from a statistical standpoint.

Whether opting for a postprocessor or an ensemble forecasting system, the following requirements should be met:

- Availability of a calibrated hydrological model.
- Access to a numerical weather prediction system that provides future inputs to the hydrological model (e.g. precipitation and temperature forecasts).
- Computational capacity to run the postprocessor or ensemble forecasting system.

In Southern Africa, and particularly in the Zambezi River Basin, some of these requirements have been hard to fulfil. While presently the Zambezi Watercourse Commission (ZAMCOM) is making concrete efforts towards the operationalization of a centralized forecasting system for the basin, until very recently major reservoirs upon whose hydropower production the regional economy depends did not have access to a probabilistic flow forecasting systems. This is perhaps not surprising, as flow forecasting systems are usually expensive and require a fair amount of expertise to set up and operate.

In 2016, the largest artificial reservoirs in the Zambezi River Basin, Kariba and Cahora Bassa, operated at low water levels. This affected Kariba particularly, with repercussions on the dam's hydropower production capacity and, consequently, on the regional electricity supply.

In the present contribution, a free and open source flow forecasting system designed for Africa is presented. It is a data-driven approach based on state-of-the-art machine learning models that fully adapts to the information capable of producing probabilistic forecasts.

1.2 Development of the forecasting system

The system started being developed with the African Dams Project (ADAPT) which, focusing on integrated water resources management in the Zambezi River Basin, has been a fruitful endeavor to collect and interpret data in order to increase the scientific basis for decision making [*Mertens et al.*, 2013]. Among others, the project's research partners included the Integrated Water Resources Management Centre at the University of Zambia, the Centre for Engineering Studies of the Eduardo Mondlane university (Mozambique), the hydropower operators Zesco and ZRA, the Laboratory of Hydraulic Constructions (LCH) of the École Polytechnique Fédérale de Lausanne (EPFL, Switzerland), the Department of Surface Water – Research and Management of the Swiss Federal Institute for Aquatic Science and Technology, the institutes of Integrative Biology, Environmental Engineering, Environmental Decisions, Biogeochemistry and Pollutant Dynamics, the Center for Comparative and International Studies, and Advanced Studies in Development and Cooperation at the Swiss Federal Institute of Technology in Zurich [*Mertens*, 2013].

Following ADAPT, the smaller-scale ADAPT-Database (ADAPT-DB) endeavor aimed to produce a water resources database with data analyses capabilities. It was within ADAPT-DB, which finished in 2016, that the present forecasting system was mainly developed. Working with local stakeholders, the project benefitted particularly from the engagement and interest of the Water Resources Institute of Angola (*Instituto Nacional de Recursos Hídricos*), the Zambezi River Authority (managing the Kariba dam), the Eduardo Mondlane University in Mozambique (*Universidade Eduardo Mondlane*), and the Hydroelectric of Cahora Bassa (*Hidroeléctrica de Cahora Bassa*, managing the Cahora Bassa dam).

At its core, the forecasting system developed under ADAPT-DB can predict an inverse conditional distribution based on a sufficiently long historical series of observed values and ancillary data. For a given probability, p, the value of the conditional distribution of a random variable Y conditioned on the set of ancillary data D, $\hat{F}^{-1}_{Y|D}(p)$, can be estimated according to Eq. (1):

$$\hat{F}^{-1}_{Y|D}(p) = \sum_{i=1}^{n} \frac{g_i(\mathbf{d} \mid \theta_i)}{n}, \quad \forall g_i : |\eta(g_i) - p| < \beta \tag{1}$$

where $g_i(\cdot)$ represents a chosen deterministic model (an artificial neural network in this case [*Haykin*, 1994; *ASCE*, 2000a, 2000b]) n is the size on an ensemble of models used to make the prediction, $\mathbf{d}$ is the data the prediction is conditioned upon, θ_i are the parameters of model $g_i(\cdot)$, $\eta(g_i)$ is the fraction of observations that exceed the simulations made by $g_i(\cdot)$, and β is a small number that allows to base the prediction on models which are in line with the probability p.

Finding the parameters θ_i for every deterministic model in the ensemble is the computationally challenging part of the methodology. That task is accomplished by employing a multi-objective optimization algorithm [*Coello Coello and Lechuga*, 2002; *Deb et al.*, 2002] to find the Pareto surface spanned by exceedance, $\eta(g_i)$, and a chosen error metric that ranks models according to how well their simulations match observations (for example, the mean absolute error).

Owing to its general nature, the concept can in principle be applied to model any stationary phenomena.

2. Case study

In the scope of this work the flow forecasting system will be applied to the Upper Zambezi River Basin (Fig. 1). The Upper Zambezi is a large catchment with an area of 519 000 km^2. It contributes with approximately 80% of the inflows to the Kariba reservoir, which is the largest artificial reservoir in the World with a total volume of about 180 km^3. More details on the catchment and the Kariba hydropower system can be found in Matos [2014] and Schleiss and Matos [2016].

The concentration time of the Upper Zambezi has been estimated to be of about 60 days [*Matos*, 2014]. Along the main stem of the Zambezi, there are several locations where discharges and water stages are recorded daily. The ones considered in this work, from upstream to downstream, are Chavuma Falls, Lukulu, Senanga, and Victoria Falls. A major wetland is present in the catchment: the Barotse Flats. Due to its massive storage capacity, it has a significant effect on the hydrograph of the Zambezi between Lukulu and Senanga.

To illustrate the capabilities of the proposed forecasting system, flow and volume predictions are made at Victoria Falls based on different sets of information.

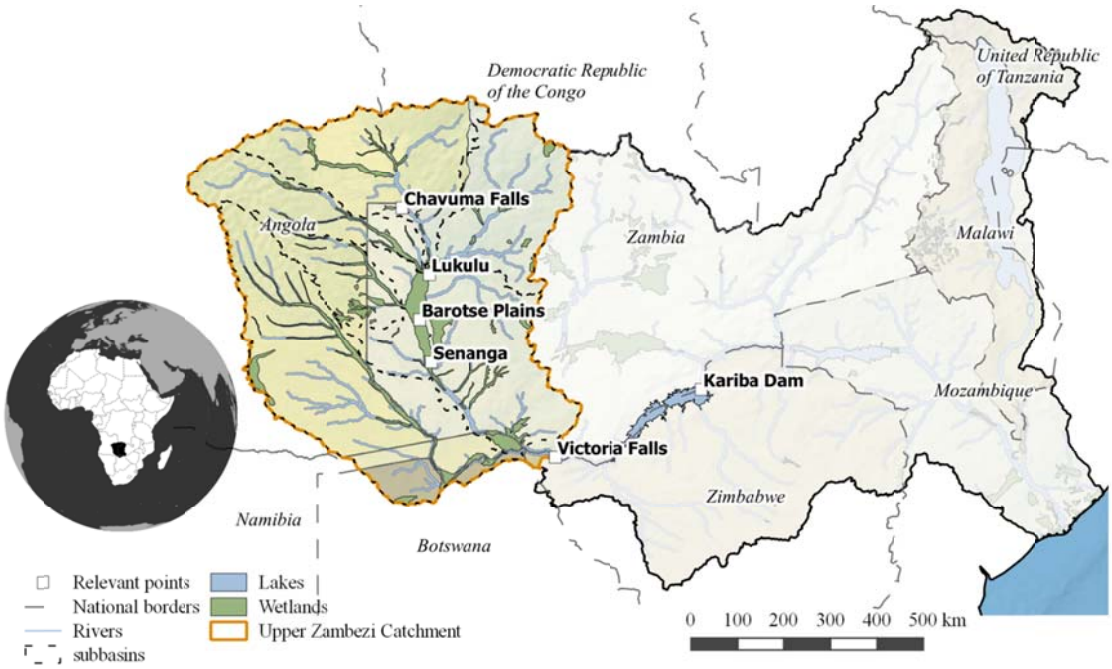

Fig. 1. Upper Zambezi River Basin and its main features of interest.

3. Interpreting probabilistic forecasts

The interpretation of probabilistic forecasts is somewhat different from the interpretation of deterministic ones. In fact, because the output of a probabilistic forecast is a predictive uncertainty distribution instead of a single value, its quality can be viewed in terms of reliability and resolution.

The resolution of a model can be well quantified resorting to the α metric, described in *Renard et al.* [2010]. Defined by Eqs. (2) and (3), it reflects the overall reliability of the estimated probability distribution. α values vary between 0 and 1, the latter indicating perfect reliability.

$$\alpha = 1 - 2\alpha' \tag{2}$$

$$\alpha' = \sum_{i=1}^{N} \left| p_{(i)} - p_{(i)}^{th} \right| \Big/ N \tag{3}$$

In Eq. (3), $p_{(i)}$ and $p_{(i)}^{th}$ stand for the i^{th} observed and theoretical p-values of the prediction and *N* represents the number of predictions.

Beyond the the α metric, reliability can be evaluated resorting to a predictive QQ-plot. The interpretation of the QQ-plot is illustrated in Fig. 2.

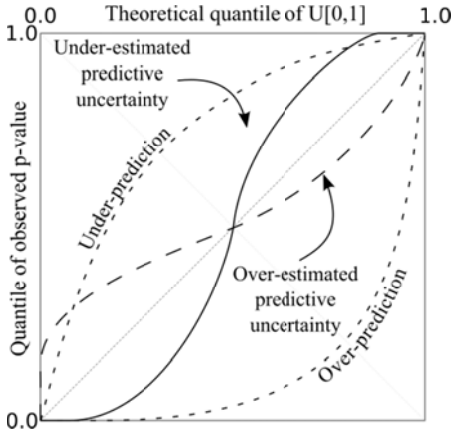

Fig. 2. Interpretation of a predictive QQ-plot. Adapted from Renard et al. [2010].

The resolution (or sharpness) of the forecast is a measure of how wide the predictive uncertainty distribution is. It can be calculated using Eq. (4) [*Renard et al.*, 2010]:

$$\pi = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{\sigma_i} \tag{4}$$

where π is the resolution, N is the number of predictions, and σ_i is the standard deviation of the predictive uncertainty distribution estimated for the i^{th} prediction. π varies between 0 and infinity and larger values are preferable.

To some extent, resolution and reliability can be independent. An example of two reliable forecasts with very different resolutions is presented in Figure 3.

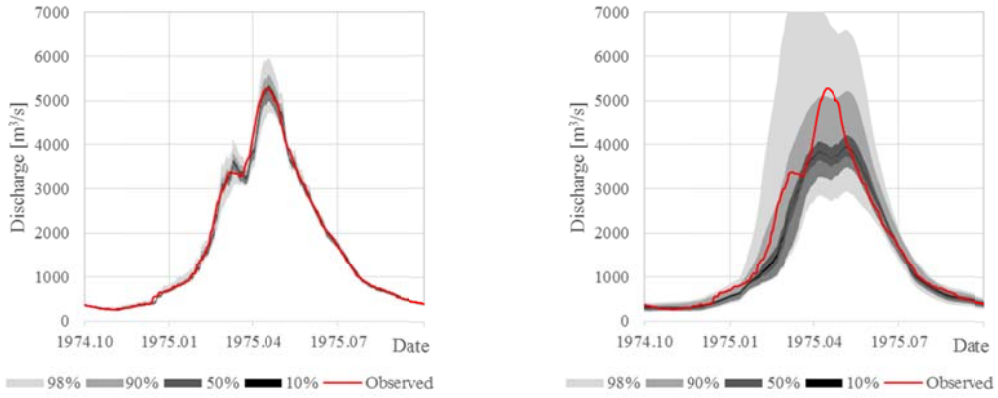

Fig. 3. Reliable forecasts of different resolutions. The percentage indicates the probability of an observation being found within each color band.

4. Results and discussion

4.1 Autoregressive forecasts

The first illustration of the system is done by carrying out a hindcast (forecast carried out on past data that can therefore be compared to historical observations) of daily discharges at Victoria Falls taking observations at that station and time of the year as the only input variables. This corresponds to the most basic form of forecast and can be accomplished whenever a long series of historical records is available for training the models used for the forecast. Results for a window of 10 years, from 2000 to 2010, are presented in Figure 4.

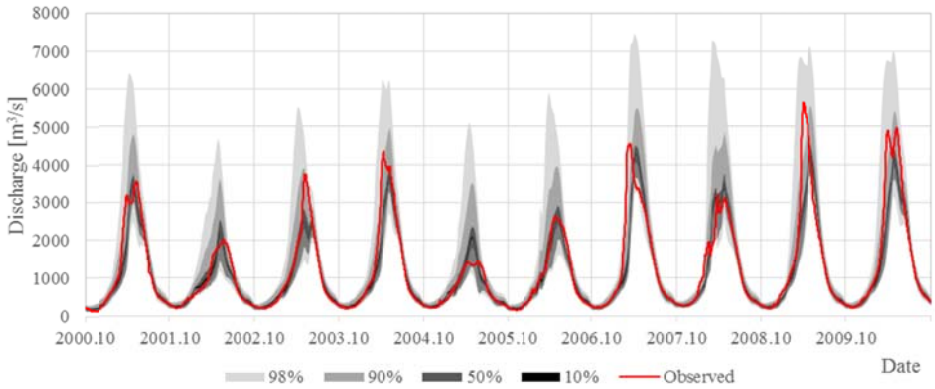

Fig. 4. Autoregressive 30-day forecast at Victoria Falls. The percentage indicates the probability of an observation being found within each color band.

A detail of the 2009-2010 hydrological year is presented in Figure 5, along with the predictive QQ-plot obtained in validation years (years not used for model training). The validation numerical performance metrics were α=0.985, and π=1.026. Evidently, a 30-day lead time forecast with this quality resorting to autoregressive data alone is only possible due to the very long concentration time of the Upper Zambezi and the smooth behavior that characterizes its hydrograph.

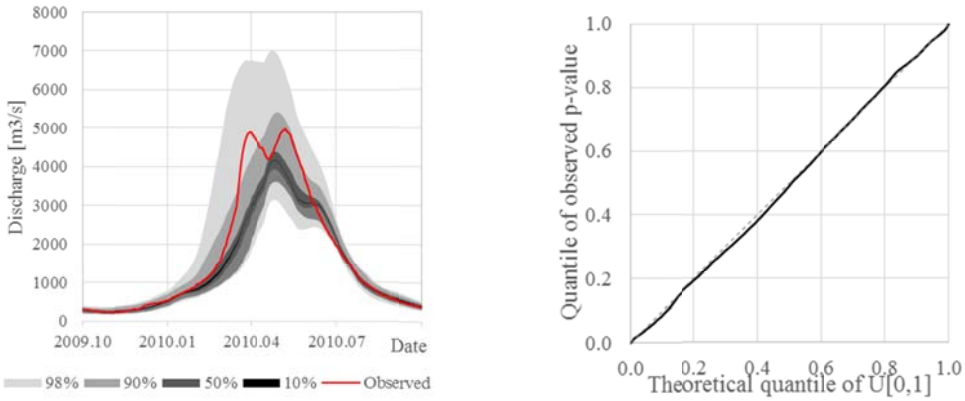

Fig. 5. Autoregressive 30-day forecast at Victoria Falls. On the left: detail of the hydrological year of 2009-2010. On the right: predictive QQ-plot for validation years.

4.2 Composed forecasts

One can naturally improve the autoregressive forecast by providing the system with additional information. This information can in theory be anything, but pertinent examples for flow forecasting are measurements from upstream stations, recorded rainfall, rainfall forecasts, or even simulations from other hydrological models.

In the present case three additional data series were used: discharges measured at Chavuma Falls, water stages recorded at Lukulu, and water stages recorded at Senanga. Without any specific pre-processing, that information was used to enhance the autoregressive 30-day lead time forecasts at Victoria Falls, as can be inferred from Figure 6. Validation performances were now α=0.916 and π=1.066 and, as can be seen, the width of the predictive distribution was significantly reduced. In Figure 7 the detail of the 2009-2010 hydrological year, as well as the validation predictive QQ-plot are presented. Through the latter, it becomes apparent that the improvement in resolution came at the expense of a small tendency for over-prediction.

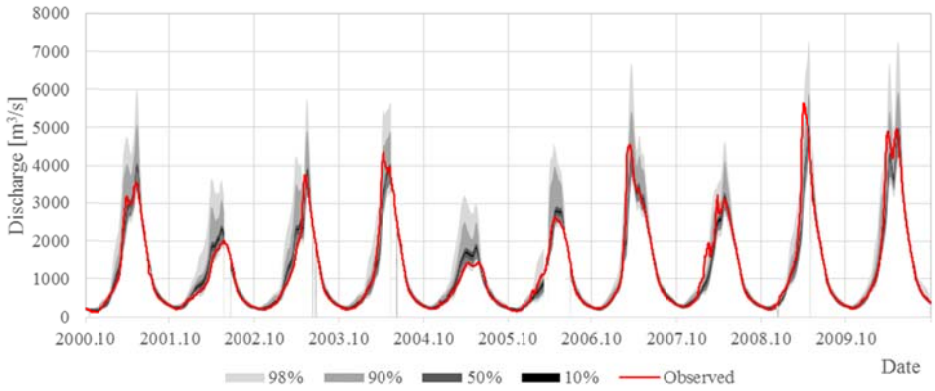

Fig. 6. Composed 30-day forecast at Victoria Falls. The percentage indicates the probability of an observation being found within each color band.

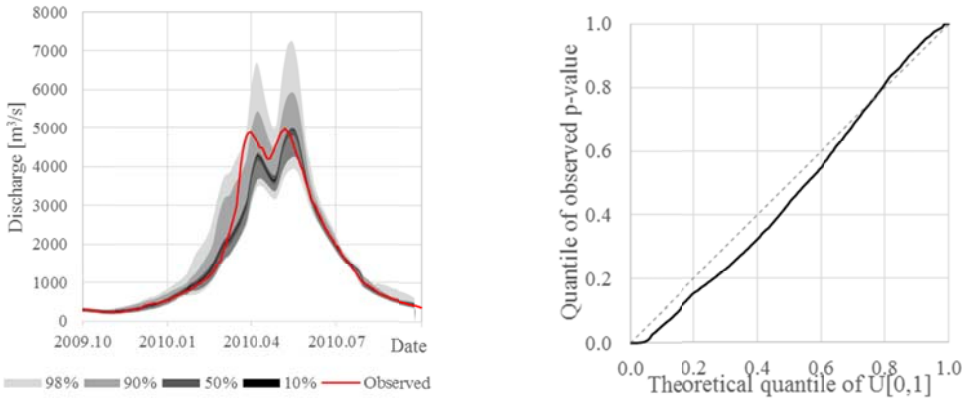

Fig. 7. Composed 30-day forecast at Victoria Falls. On the left: detail of the hydrological year of 2009-2010. On the right: predictive QQ-plot for validation years.

4.3 Volume forecasts

Just as arbitrary types of data can be used as inputs, the object of the forecasts is not limited to flows. I can also be, for example, water stage or accumulated volume. As the latter has a particular interest for reservoir management, the last case presented in this paper is a forecast of the volume of water that will pass in Victoria Falls in the 90 days following the date of reference. A hindcast of the prediction, from 2000 to 2010, is depicted in Figure 8.

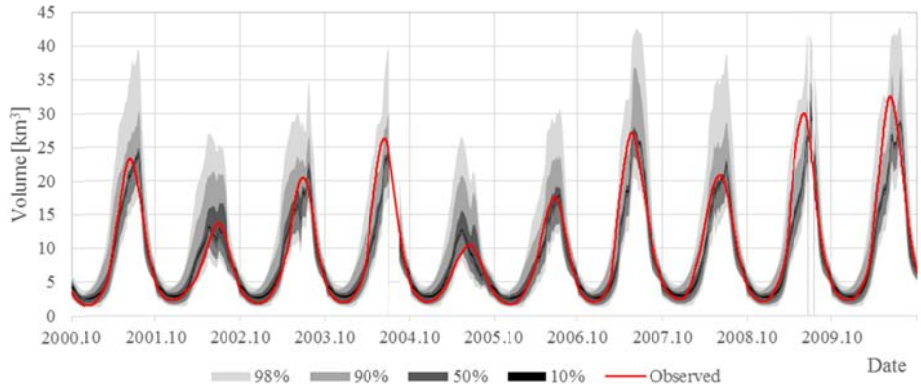

Fig. 8. 90-day accumulated volume forecasts at Victoria Falls. The percentage indicates the probability of an observation being found within each color band.

The numerical validation metrics were α=0.959 and π=1.014. In Figure 9, the detail of the hydrological year of 2009-2010 and the validation predictive QQ-plot are presented. It is worth noticing that, at long lead times, probabilistic predictions become particularly useful. In fact, forecasting 90-days into the future is hard – even in the Upper Zambezi – and deterministic predictions will always miss observations. Probabilistic predictions, however, acknowledge that errors are likely to be made and supply the modeler with reliable information.

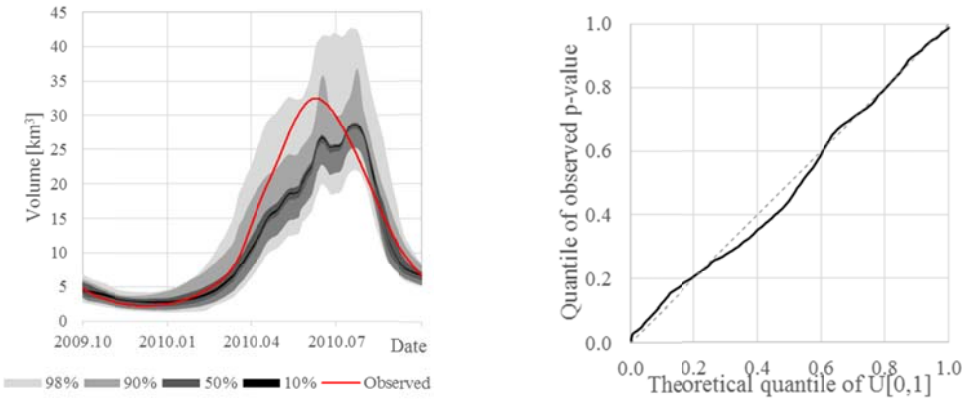

Fig. 9. 90-day accumulated volume forecasts at Victoria Falls. On the left: detail of the hydrological year of 2009-2010. On the right: predictive QQ-plot for validation years.

5. Conclusions

The flow forecasting tool presented in this paper is free and can run on a normal desktop computer. It is presently under active development and being tested jointly with stakeholders in the Zambezi River Basin.

More than a flow forecasting methodology, this tool also takes the role of an online database that can be accessed remotely through a common internet browser such as Mozilla Firefox or Google Chrome without the need to install any specific software. It also does not require programming skills to be used. Through its flexible principle, the forecasting system can predict flows, water stages or even accumulated volumes at varying lead times, always making good use of the information that is available.

The methodology relies on a data-driven principle. As such, a representative historical data series is extremely important to obtain reliable forecasts. Also, the series being modelled should be stationary and, therefore, drifts such as the climate change cannot be directly handled. Finally, moving towards extreme events, the estimates of predictive uncertainty distribution lose quality and should not be expected to be reliable.

The proposed forecasting tool is free, easy to use, and highly flexible. It can be retrieved at http://zambezi.epfl.ch. In cases where other forecasting methodologies are too expensive to set up, too complex to prepare and maintain, or simply not leading to reliable results, it can constitute a powerful and practical alternative way of producing hydrological forecasts.

Acknowledgments

The authors would like to acknowledge the Zambezi River Authority, ZESCO, and the Department of Water Affairs of Zambia, who collected and provided the data needed to undertake this study. Also, the authors thank the Swiss Competence Center Environment and Sustainability (CCES) for having provided the funding needed to accomplish this project.

References

1. **ASCE (a)**, "Artificial neural network in hydrology. II: hydrologic applications", *J. Hydrol*. Eng., 5(2), 124–137, doi:10.1061/(ASCE)1084-0699(2000)5:2(124), 2000.
2. **ASCE (b)**, "Artificial neural networks in hydrology. I: preliminary concepts", *J. Hydrol*. Eng., 5(2), 115–123, doi:10.1061/(ASCE)1084-0699(2000)5:2(115), 2000.
3. **Cloke, H. L., and F. Pappenberger**, "Ensemble flood forecasting: A review", *J. Hydrol.*, 375(3–4), 613–626, doi:10.1016/j.jhydrol.2009.06.005, 2009.
4. **Coello Coello, C. A., and M. S. Lechuga**, "MOPSO: A proposal for multiple objective particle swarm optimization", *Proceedings*, 2002 Congress on Evolutionary Computation, CEC 2002, vol. 2, pp. 1051–1056, 2002.
5. **Deb, K., A. Pratap, S. Agarwal, and T. Meyarivan**, "A fast and elitist multiobjective genetic algorithm: NSGA-II", *IEEE Trans. Evol. Comput*., 6(2), 182–197, doi:10.1109/4235.996017, 2002.
6. **Haykin, S.**, "Neural networks: a comprehensive foundation", Prentice Hall, Upper Saddle River, NJ, USA, 1994.
7. **Krzysztofowicz, R.**, "Bayesian system for probabilistic river stage forecasting", *J. Hydrol.*, 268(1–4), 16–40, doi:10.1016/S0022-1694(02)00106-3, 2002.

8. **Matos, J. P.**, "Hydraulic-hydrologic model for the Zambezi River using satellite data and artificial intelligence techniques", *Ph.D. Thesis No. 6225*, École Polytechnique Fédérale de Lausanne and the University of Lisbon, Lausanne, Switzerland, 2014 and *Communication 60 of the Laboratory of Hydraulic Constructions of the École Polytechnique Fédérale de Lausanne*, edited by A. J. Schleiss, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2014.
9. **Mertens, J.**, "African Dams Project: an integrated water resources management study". *Final stakeholder report*, ETH Zurich, Zurich, 2013.
10. **Mertens, J., B. Wehrli, A. Tilmant, A. Schleiss, T. Cohen Liechti, and J. P. Matos**, "Adapted reservoir management in the Zambezi river basin to meet environmental needs", *Int. J. Hydropower Dams*, 20(2), 80–84, 2013.
11. **Renard, B., D. Kavetski, G. Kuczera, M. Thyer, and S. W. Franks**, "Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors", *Water Resour. Res.*, 46(5), W05521, doi:10.1029/2009WR008328, 2010.
12. **Schleiss, A. J., and J. P. Matos**, "Zambezi River Basin", *in Chow's Handbook of Applied Hydrology*, edited by V. P. Singh, 98-1–98-6, McGraw-Hill Education, 2016.
13. **Solomatine, D. P., and D. L. Shrestha**, "A novel method to estimate model uncertainty using machine learning techniques", *Water Resour. Res.*, 45(12), n/a-n/a, doi:10.1029/2008WR006839, 2009.
14. **Todini, E.**, "A model conditional processor to assess predictive uncertainty in flood forecasting", *Int. J. River Basin Manag.*, 6(2), 123–137, doi:10.1080/15715124.2008.9635342, 2008.
15. **Weerts, A. H., H. C. Winsemius, and J. S. Verkade**, "Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)", *Hydrol. Earth Syst. Sci.*, 15(1), 255–265, doi:10.5194/hess-15-255-2011, 2011.

The Authors

J.P. Matos is a Postdoctoral Researcher at the Laboratory of Hydraulic Constructions of the École Polytechnique Fédérale de Lausanne (EPFL). His PhD focused on the development of a hydrologic model for the Zambezi River basin using satellite data and artificial intelligence techniques. He holds an MSc in Civil Engineering from the Technical University of Lisbon and has worked as a consultant in hydrology, water supply, and sanitation. He is interested in the fields of risk assessment, machine learning, remote sensing, and optimization of complex non-linear systems.

Prof Dr A.J. Schleiss graduated in Civil Engineering from the Swiss Federal Institute of Technology (ETH) in Zurich, Switzerland, in 1978. After joining the Laboratory of Hydraulic, Hydrology and Glaciology at ETH as a research associate and senior assistant, he obtained a Doctorate of Technical Sciences on the topic of pressure tunnel design in 1986. He then worked for 11 years for Electrowatt Engineering Ltd in Zurich, and was involved in the design of many hydro projects around the world as an expert on hydraulic engineering and underground waterways. In 1997 he was nominated full professor and became Director of the Laboratory of Hydraulic Constructions (LCH) at the Swiss Federal Institute of Technology Lausanne (EPFL). LCH's activities comprise education, research and services in the field of both fundamental and applied hydraulics and design of hydraulic structures and schemes. The research focuses on the interaction between water, sediment-rock, air and hydraulic structures as well as associated environmental issues and involves both numerical and physical modeling. Prof. Schleiss is also involved as an international expert on several dam and hydropower projects throughout the world. From 2006 to 2012 he was Director of the Civil Engineering program of EPFL and Chairman of the Swiss Committee on Dams. In 2006 he obtained the ASCE Karl Emil Hilgard Hydraulic Prize as well as the J. C. Stevens Award. He was listed in 2011 among the 20 international personalities that "have made the biggest difference to the sector of Water Power & Dam Construction over the last 10 years". For his outstanding contributions to advance the art and science of hydraulic structures engineering he obtained in 2015 the ASCE-EWRI Hydraulic Structures Medal. After having served as vice-president between 2012 and 2015 he was elected president of the International Commission on Large Dams (ICOLD) in 2015.