# Book of Short Papers SIS 2018

*Editors:* Antonino Abbruzzo - Eugenio Brentari

Marcello Chiodi - Davide Piacentino

SIS 2018

49 TH SCIENTIFIC MEETING
OF THE ITALIAN
STATISTICAL SOCIETY

PALERMO 20–22 JUNE

PUBLISHED BY PEARSON

WWW.PEARSON.COM

*First printing, novembre 2018 - ISBN-9788891910233*

# Contents

# Approximate Bayesian Computation for Forecasting in Hydrological models

## *Metodi Bayesian approssimati per le previsioni nei modelli idrologici*

Jonathan Romero-Cuéllar, Antonino Abbruzzo, Giada Adelfio and Félix Francés

**Abstract** Approximate Bayesian Computation (ABC) is a statistical tool for handling parameter inference in a range of challenging statistical problems, mostly characterized by an intractable likelihood function. In this paper, we focus on the application of ABC to hydrological models, not as a tool for parametric inference, but as a mechanism for generating probabilistic forecasts. This mechanism is referred as Approximate Bayesian Forecasting (ABF). The abcd water balance model is applied to a case study on Aipe river basin in Columbia to demonstrate the applicability of ABF. The predictivity of the ABF is compared with the predictivity of the MCMC algorithm. The results show that the ABF method as similar performance as the MCMC algorithm in terms of forecasting. Despite the latter is a very flexible tool and it usually gives better parameter estimates it needs a tractable likelihood.

**Abstract** *In questo articolo, il metodo chiamato Approximate Bayesian Computation (ABC) viene applicato ai modelli idrologici, non come uno strumento per l'inferenza parametrica, ma come un meccanismo per generare previsioni probabilistiche, dando luogo all'Approximate Bayesian Forecasting (ABF). L'ABF è applicato a un caso studio sul bacino del fiume Aipe in Colombia. Viene considerato un modello idrologico semplice per dimostrare l'applicabilità di ABF e confrontarlo con la predittività del metodo MCMC. Nonostante i risultati mostrano che il metodo ABF e l'algoritmo MCMC non differiscono in termmini di previsioni ottenute, l'ABF è comunque uno strumento molto flessibile e fornisce risultati utili anche quando si è in presenza di una verosimiglianza intrattabile.*

**Key words:** Predictive uncertainty, Probabilistic post-processing approach, Bayesian forecasting, Sufficient statistics, Hydrological models, Intractable likelihood.

---

Jonathan Romero-Cuéllar and Félix Francés
Research Institute on Water and Environmental Engineering, Universitat Politécnica de Valéncia, Spain, e-mail: jorocue1@doctor.upv.es

Antonino Abbruzzo and Giada Adelfio
Department of Economics, Business and Statistical Sciences, Universitá degli Studi di Palermo, Italia, e-mail: antonino.abbruzzo@unipa.it

# 1 Introduction

In hydrological models, predictions are crucial for supporting decision-making and water management. Reliability of prediction of hydrologic outcomes is affected by several sources of uncertainty such as input or forcing data uncertainty, initial conditions, model uncertainty or epistemic error, parameters inference, output uncertainty. So several sources of uncertainty affect the full predictive uncertainty, that is the probability of occurrence of a future value of a response variable (streamflow, water level) conditional on all the covariates, usually provided by forecasting models [9]. Therefore, the forecast approaches, rather than looking for deterministic predictions, essentially aim at quantifying predictive uncertainty. Predictive uncertainty estimation in hydrological models is a challenge when dealing with intractable likelihoods. The Approximate Bayesian Computation (ABC) overcomes the likelihood-based approach via the use of sufficient statistics and simulated data [1]. The idea behind the ABC approach was first introduced in population and evolutionary genetics [6, 7]. The ABC has a wide range of application domains because it is useful when an explicit likelihood function cannot be justified [10]. The main focus, of the most of the studies about the ABC, is the quantification of uncertainty about parameters [2]. Together with the increasing applications of ABC (see [4] for recent surveys), attention has recently been paid to the theoretical properties of the method, including the asymptotic behaviour of: the ABC posterior distributions, the point estimates derived from those distributions, and the Bayes factors that condition on summaries (see for instance [4]). The ABC approach in hydrological models is introduced in [11], using the ABC to estimate posterior distributions of parameters for simulation-based models.

The aim of this paper is to introduce the ABC as an approach for generating probabilistic forecasts in hydrological models. This approach is referred to Approximate Bayesian Forecasting (ABF) [2]. A streamflow forecasting on a case study of the Aipe river basin in Columbia is used to show the potential strength of the ABF. Predictions derived from the ABF algorithm are compared to prediction derived from the MCMC algorithm.

This paper is structured as follows. In the first section, we describe a simple hydrological model. In the second section, we describe the application of the ABC for the hydrological model. In the third section, we compare the ABF and the MCMC algorithm.

# 2 Approxiamte Bayesian Forecasting for the hydrological model

The abcd water balance model is a hydrological model for simulating streamflow (see [8]). This model is a fairly general conceptual rainfall-runoff model which transforms rainfall and potential evapotranspiration data to streamflow at the catchment outlet. The model is comprised of two storage compartments: soil moisture and groundwater. The soil moisture gains water from precipitation and loses water to evapotranspiration, surface runoff, and groundwater recharge. The groundwater compartment gains water from recharge and loses water as discharge. The total streamflow, which is the outcome we are interested in, is the sum of surface runoff from the soil moisture and groundwater discharge. It applies the continuity equation to a control volume representing the upper soil zone, from which evapotranspiration is assumed to occur, so that

$$Sw_t + ET_t + Q_t + R_t = Sw_{t-1} + P_t, \tag{1}$$

where $P_t$ is a total precipitation for the month, $ET_t$ is actual evapotranspiration, $R_t$ is recharge to groundwater storage, $Q_t$ is upper zone contribution to runoff, and $Sw_t$ and $Sw_{t-1}$ represent upper soil zone soil moisture storage at the current and previous time steps respectively. For the groundwater component, the mass balance equation is

$$Sg_t + Qg_t = Sg_{t-1} + R_t, \tag{2}$$

where $Qg_t$ is groundwater discharge, $Sg_t$ and $Sg_{t-1}$ represent groundwater storage at the current and previous time steps, respectively. More details about the abcd water balance model are in [8] and [5]. Equations (1) and (2) produce the streamflow output for $t$ times. We denote this variable by $\tilde{\mathbf{y}} = \{\tilde{y}_1, \ldots, \tilde{y}_t, \ldots, \tilde{y}_T\}$. Starting from this result, the ABC is used as hydrologic post-processor. Hydrologic post-processing works directly on hydrologic model outputs by using a statistical model to represent the relationship between model outputs and corresponding observations. It serves the purpose of removing model biases from all upstream uncertainty sources. In this paper, we use the ABC to estimate the parameters of the linear model

$$y_t = \beta_0 + \beta_1 \tilde{y}_t + \varepsilon_t, \tag{3}$$

where $y_t$ is the observed streamflow at time $t$, $\beta_0$ and $\beta_1$ are parameters, $\tilde{y}_t$ is the output form the abcd model (equations (1) and (2)). The random variable $\varepsilon_t$ is the error term in the model, representing random fluctuations, i.e. the effect of factor outside of our control or measurement, such that $\varepsilon_t \sim N(0, \sigma^2)$, i.i.d. Specifically, the ABC produces draws from an approximation of the posterior distribution of $\theta = (\beta_0, \beta_1, \sigma^2)$, i.e.

$$p(\theta|\mathbf{y}) \propto p(\mathbf{y}|\theta)p(\theta),$$

where $p(\theta)$ is the prior, $p(\mathbf{y}|\theta)$ is the distribution of $\mathbf{y}$ conditional on the parameters. Even though we can use ABC with intractable likelihood $\ell(\theta|\mathbf{y})$, we must be able to simulate data from $p(\theta)$ and $p(\mathbf{y}|\theta)$. We assume flat normal priors for $\beta_0$ and $\beta_1$ and $Y_t|\theta \sim N(\mu_t = \beta_0 + \beta_1 \tilde{y}_t, \sigma^2)$. The pseudo code for the ABC is summarized in Algorithm 1. Algorithm 1 thus samples $\theta$ and pseudo-data $\mathbf{z}$ from the joint posterior:

$$p_\varepsilon(\theta, \mathbf{z}|\eta(\mathbf{y})) = \frac{p(\theta)p(\mathbf{z}|\theta)\mathbb{I}(\mathbf{z})}{\int_\Theta \int_\mathbf{z} p(\theta)p(\mathbf{z}|\theta)\mathbb{I}(\mathbf{z})d\Theta d\mathbf{z}} \tag{4}$$

The ABF produces the approximate predictive uncentanty formally defined as

**Algorithm 1** ABC accept/reject algorithm

1: $\theta^i$, $i = 1,\dots,N$ from $p(\theta)$
2: $\mathbf{z}^i = (z_1^i, z_2^i, \dots, z_T^i)^\top$, $i = 1,\dots,N$, from the likelihood, $p(\cdot|\theta^i)$
3: Select $\theta^i$ such that:

$$d\{\eta(\mathbf{y}), \eta(\mathbf{z}^i)\} \leq \varepsilon$$

where $\eta(\cdot)$ is a vector statistic, $d\{\cdot\}$ is a distance criterion, and, given $N$, the tolarence level $\varepsilon$ is chosen to be small.

$$g(y_{T+1}|\mathbf{y}) = \int_\Theta p(y_{T+1}|\theta, \mathbf{y}) p_\varepsilon(\theta|\eta(\mathbf{y})) d\theta, \tag{5}$$

where $p_\varepsilon(\theta|\eta(\mathbf{y})) = \int_\mathbf{z} p_\varepsilon(\theta, \mathbf{z}|\eta(\mathbf{y})) d\mathbf{z}$.

## 3 Streamflow data analysis

We use monthly data of mean areal precipitation, mean areal potential evaporation, and so the other variables of the abcd model, from the Aipe river basin at Huila, Colombia, that is a tropical basin described in the study by [3]. Fig. 1 represents some characteristics of the hydrological behavior of the Aipe river basin. We ap-

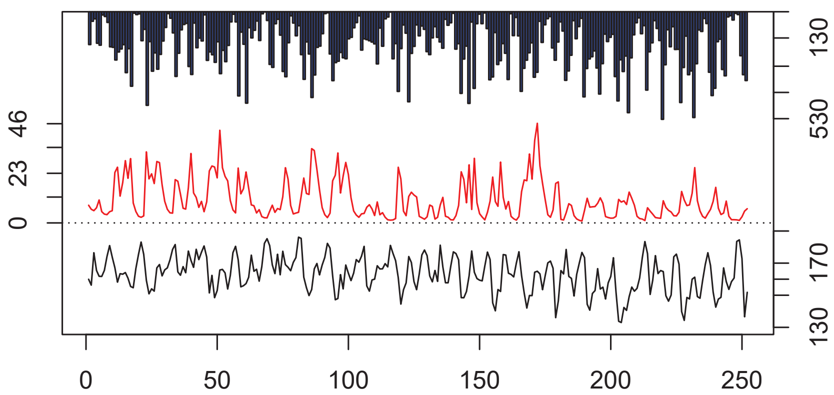


**Fig. 1** Monthly time series of Aipe river basin. The blue histogram corresponds to the rainfall (mm), the red line corresponds to the runoff $(m^3 s^{-1})$ and the black line corresponds to the potential evaporation (mm).

ply both the ABC and the MCMC algorithm to obtain parameters estimation of the model (3). Moreover, the ABF is used to assess the predictive uncertainty and compared to the MCMC predictive distribution. The MCMC algorithm is used as a benchmark, since it takes advantage of the likelihood which, for the model we are dealing with, is tractable. To produce the results for the ABF we set the Euclidean distance and choose the mean and the standard deviation as sufficient statistics. In Fig. 2 we show the results of the MCMC and ABC approaches. Although the MCMC and ABC posteriors for both the elements of $\theta = (\beta_0, \beta_1)$ are quite a different one from the other (panel on the left and in the middle), the predictive distributions are quite close and similar.

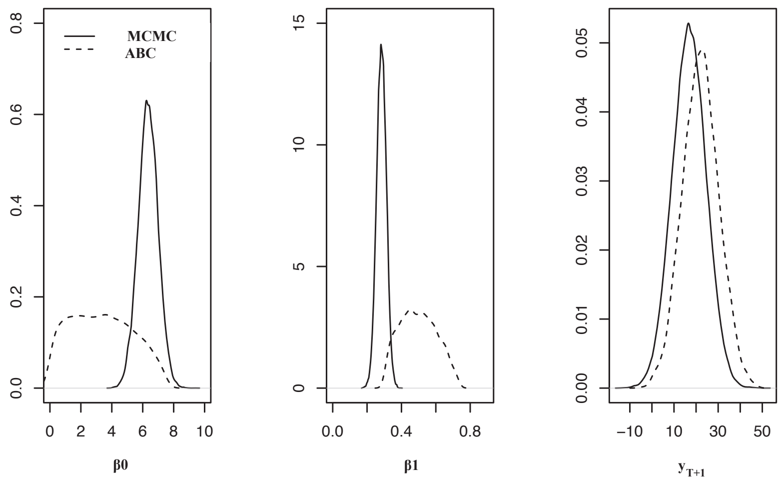


**Fig. 2** Marginal posteriors fro the parameters, both considering the MCMC and ABC methods for the monthly time series of Aipe river basin (on the left and in the middle). Predictive density functions both MCMC and ABC (on the right).

## 4 Conclusion

In this paper, we discuss the use of the Approximate Bayesian Forecasting for hydrological models. The advantage of this tech niche is the applicability for intractable likelihood. This characteristic can make this model very appealing in hydrological forecasting. Even though the ABC seems inappropriate for parameter estimation (probably due to the choice of sufficient statistics) it shows good performance (similar to the MCMC algorithm) in terms of prediction.

## References

[1] Csilléry K., François O. and Blum M.G.B. Abc: An R package for approximate Bayesian computation (ABC). *Methods in Ecology and Evolution*, 3(3):475–479, 2012.

[2] Frazier D. T., Maneesoonthorn W., Martin G. M. , and McCabe B. P.M.. Approximate bayesian forecasting. *arXiv preprint arXiv:1712.07750*, 2017.

[3] Labrador A. F., Zúñiga J. M, and Romero J.. Development of a model for integral planning of water resources in Aipe catchment, Huila, Colombia. *Revista Ingeniería y Región.*, 15(1):23–35, 2016.

[4] Marin J.M., Pudlo P., Robert C.P., and Ryder R.J.. Approximate bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

[5] Martinez G. F. and Gupta H. V.. Toward improved identification of hydrological models: A diagnostic evaluation of the "abcd" monthly water balance model for the conterminous united states. *Water Resources Research*, 46(8): W08507, 2010.

[6] Pritchard J. K., Seielstad M. T., Perez-Lezaun A., and Feldman M. W. Population growth of human y chromosomes: a study of y chromosome microsatellites. *Molecular biology and evolution*, 16(12):1791–1798, 1999.

[7] Tavaré S., Balding D. J, Griffiths R. C, and Donnelly P.. Inferring coalescence times from dna sequence data. *Genetics*, 145(2):505–518, 1997.

[8] Harold Thomas. Improved methods for national water assessment, water resources contract: WR15249270. Technical report, Harvard University, Cambridge, 1981.

[9] Todini E.. A model conditional processor to assess predictive uncertainty in flood forecasting. *International Journal of River Basin Management*, 6(2): 123–137, jun 2008.

[10] Turner B. M. and Van Zandt T.. A tutorial on approximate Bayesian computation. *Journal of Mathematical Psychology*, 56(2):69–85, 2012.

[11] Vrugt J. A. and Sadegh M. Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resources Research*, 49: 4335–4345, 2013.