02-023

K NEAREST NEIRGHBOR ALGORITHM FOR MODELLING RAINFALL HYETOGRAPHS WITH INTEREST ON HYDROLOGY

Zubelzu Mínguez, Sergio (1); Matendo, Sara (1); Galán, Victor (1); Zanella, Andrea (2); Bennis, Mehdi (3)

(1) Universidad Politécnica de Madrid, (2) University of Padova, (3) University of Oulu (2)

Practical hydrology aimed at infrastructure design, land use planning or urban planning has traditionally focused on extreme events, estimating a peak flow that should occur for a given return period. This approximation is simple but incomplete and has sometimes been justified by the difficulty of obtaining data to simulate realistic hyetograms. Currently data-driven models, together with enhanced computational capabilities, have opened up new opportunities for hydrological simulation. Non-parametric models like the KNN are simple but allow efficient capturing of non-linear models. In the present work the results of the use of this type of algorithms for the simulation of hydrographs are exposed.

Keywords: Hydrology; Modelling, Data-driven models, Machine learning; Hyetographs.

ALGORITMO K NEAREST NEIRGHBOR PARA LA SIMULACIÓN DE HIETOGRAMAS DE USO EN LOS ESTUDIOS HIDROLÓGICOS

Tradicionalmente la hidrología en la práctica del diseño de infraestructuras, ordenación del territorio o planeamiento urbanístico se ha limitado al estudio de eventos extremos estimando un caudal punta que habría de producirse para un determinado período de retorno. Esta aproximación es simple pero incompleta y se ha justificado en ocasiones en la dificultad de obtener datos para simular hietogramas realistas. En la actualidad los modelos de datos y las capacidades computacionales han abierto nuevas oportunidades para la simulación hidrológica. Los modelos no paramétricos como el KNN son simples pero permiten capturar de forma eficiente modelos no lineales. En el presente trabajo se exponen los resultados del uso de este tipo de algoritmos para la simulación de hidrogramas.

Palabras clave: Hidrología; Modelización; Algoritmos de datos, Aprendizaje Automático; Hietogramas.

Correspondencia: Sergio Zubelzu. Correo: sergio.zubelzu@upm.es

Agradecimientos: La publicación es parte del proyecto PCI2020-120694-2, financiado por CIN/AEI/10.13039/501100011033 y por la Unión Europea “NextGenerationEU”/PRTR

1. Introduction

In recent years, several attempts have been made to improve the understanding of physical processes and refining theories by means of accurate experimental measurements. Moreover, the advances in computational capabilities have been exploited for solving complex non-linear systems o partial differential equations applied to hydrology. Recent developments in computing and data-driven models have also opened up promising opportunities in the field of hydrological modelling, allowing new insights and approaches for modelling and forecasting hydrological processes.

Despite the apparent maturity of the theories and the advanced tools available for data collection and processing, and for solving hydrological theories, the complete understanding of the complex interactions among the physical processes underlying the hydrological phenomena still remain elusive (Blume et al., 2017). As a set of 230 authors, including well-known hydrologists and scientists from other hydrology-related disciplines, have recently highlighted in (Blösch et al., 2017) some relevant issues that still remain unsolved in hydrology. The authors presented a set of 23 unsolved problems including the ability of existing hydrologic laws for properly modelling processes at different scales, the need for innovative technologies for data collecting and modelling, the use of historical data vs soft data or the reduction of the amount of model structural/parameter/input uncertainty in hydrological prediction, among others.

Physical models for hydrology are coherent from a theoretical point of view but face a number of shortcomings complicating its wide use in practical applications. Physically-based models are often built upon non-linear systems of partial differential equations without general analytical solutions. A number of computational issues, complex parametrization or inability to deal with stochastic process hinder their use in practical applications. There are also a great variety of empirical models not providing efficient and accurate solutions since they lack of generality to be extensively used.

In recent times, hydrological modelling has also benefited from the enhanced performance of data-driven models in close conjunction with recent developments in computation techniques and the amount of data available (see for example Ali et al., 2020; Farajzadeh and Alizadeh, 2003; Ivkovic et al., 2018; Seed, 2003; Mueller et al., 2003; Rasmussen et al., 2003; Seed, 2004; Ryu et al., 2020, Fox and Wickle, 2005 or Ruzanski et al., 2011 among many others). Data-driven models open up interesting opportunities to deal with most of the unsolved modelling issues related to hydrology. Scientifics from different disciplines have presented their works applying data-driven algorithms to each process involved in hydrology and also initial attempts to come up with simplified models for the whole hydrological systems have also been exposed.

2. Literature review

If we pay attention to surface water hydrological system, once the precipitation is modelled hydrologists must focus on the transfer function relating precipitation with surface discharge. That is a problem of particular complexity since many different processes are involved and the proper modelling becomes almost impossible because a number of difficulties derived from the spatial heterogeneity arise. Literature has widely addressed this system either focusing on particular processes or on the whole system, mostly with black-box approaches in this case. A great variety of physical processes are involved in this system and, of course, data-driven models have also been widely used for this purpose.

On many occasions, the so-called precipitation-runoff processes have received researchers´ interest paying particular attention to flooding events. As in the previous cases, a number of different algorithms have been used. The conceptual approaches for defining inputs, outputs and expected relationships have also been diverse. For example, Bui et al. (2020) or Wang et al. (2020) both used neural networks for predicting flood susceptibility areas using different

topography and vegetation related variables while Pourghasemi et al. (2020) analysed the suitability of different metaheuristic approaches for flood mapping. Others focused not only on flooding susceptibility but on streamflow forecasting using auto-regressive methods.

This has probably been the most attractive field for researchers and the approaches can be grouped into two lines:

a) Pure autoregressive models, see for example Shabri and Suhartono (2012), Tikhamarine et al., (2019) or Zou (2020).

b) Combination of autoregressive and hydrological models, see for example Zhang et al. (2020) or Niu et al., (2020).

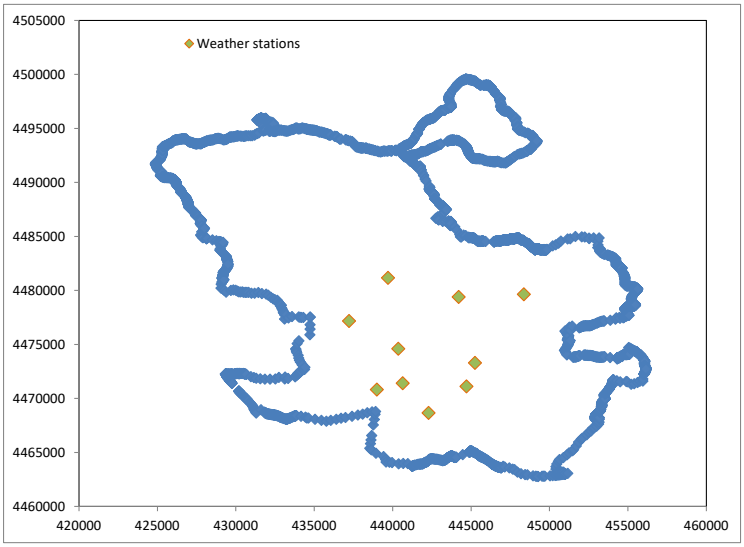
Using physical criteria to conform data-driven algorithms either for predicting, classifying or any other aim as proposed in group b) can help address individual processes in isolation as for example infiltration, runoff or transient time. This approach can be concreted not only to predict the evolution of a given target variable but also with the aim of shedding light on many complex parameters involved in physical models (soil hydraulic properties for example).

In this work we follow the second line presenting a mixed machine learning-physical approach for modelling rainfall-runoff processes merging KNN algorithms with Green-Ampt model for predicting the potential occurrence of surface runoff from rainstorm events.

3. Materials and methods

We have gathered hourly records from the weather station network belonging to Madrid City Council. The sample is compound of ten weather stations with hourly data from January 2019 to March 2021 (hereafter referred to as “102”,”103”, “106”, “107”, “108”, and “056”,) all placed inside a 30 km radius circumference. Precipitation, relative humidity and barometric pressure were available for 10 weather stations across Madrid city (see figure 1).

Figure 1. Weather stations locations



We focused on hourly precipitation records and extracted the storm events, defined as the set of precipitation records observed between two zero values (the starting time defined by a non-zero value after a zero value and the end by a non-zero value preceding a zero value).

From this operation we have then retrieved a set of empirical hyetographs with hourly data latency.

We have selected a case study, namely a catchment placed in Madrid (in the so-called Solana de Valdebebas urban planning development) and modelled the infiltration-runoff process using the Green-Ampt model (see eqs. 1 and 2).

$$F(t) = K_s t + \tau_f \Delta\theta Ln\left[1 + \frac{F(t)}{\tau_f \Delta\theta}\right] \quad (1)$$

$$f(t) = K_s\left(1 + \frac{\tau_f \Delta\theta}{F(t)}\right) \quad (2)$$

Where f(t) is the infiltration rate, F(t) is the aggregate infiltration at time *t*, Ks is the saturated hydraulic conductivity of the soil, Δθ is the difference between saturated and initial soil water contents and τ_f is the wetting front suction head, representing the suction force driving (in conjunction with gravity) the movement of the supposedly saturated wetting front. We used Neuman´s (1976) expression to estimate the wetting front suction head (eq. 3).

$$\tau_f = \frac{1}{K_s}\int_{h(\theta_i)}^{h(\theta_s)} K(\theta)\,dh(\theta) \quad (3)$$

Where h(θ) stands for the moisture-dependent water retention head and K(θ) the moisture-dependent conductivity curve.

We have thus estimated the potential runoff appearance feed a dummy variable accounting for the runoff occurrence (runoff=1) or not (runoff=0). That dummy variable is the target of the classification problem. We use the K Nearest Neighbour (KNN) algorithm to address this problem. KNN is probably the simplest machine learning algorithm. It uses information about an example´s k nearest neighbours to classify unlabelled examples. In this work we used

Euclidean distance and explored different k values taking advantage of the *train()* function of the *CARET* R library.

4. Results

4.1 Descriptive statistics

We have first analysed the recorded storms, obtaining the statistics presented in table 1.

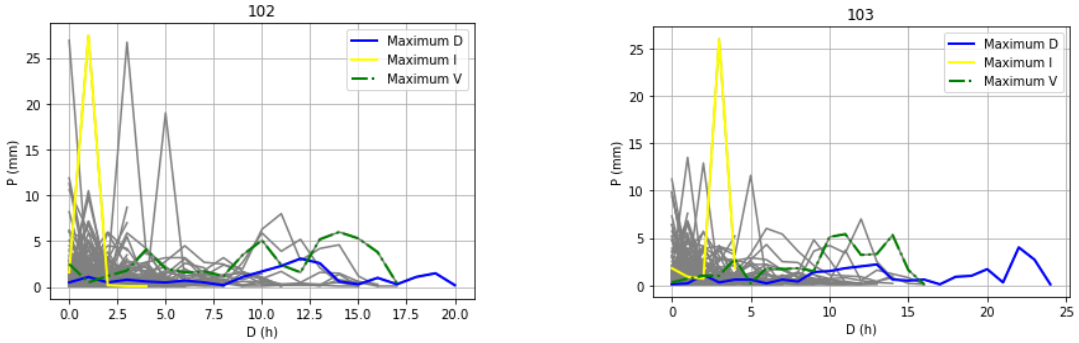
Table 1. Main statistics of the analysed storms

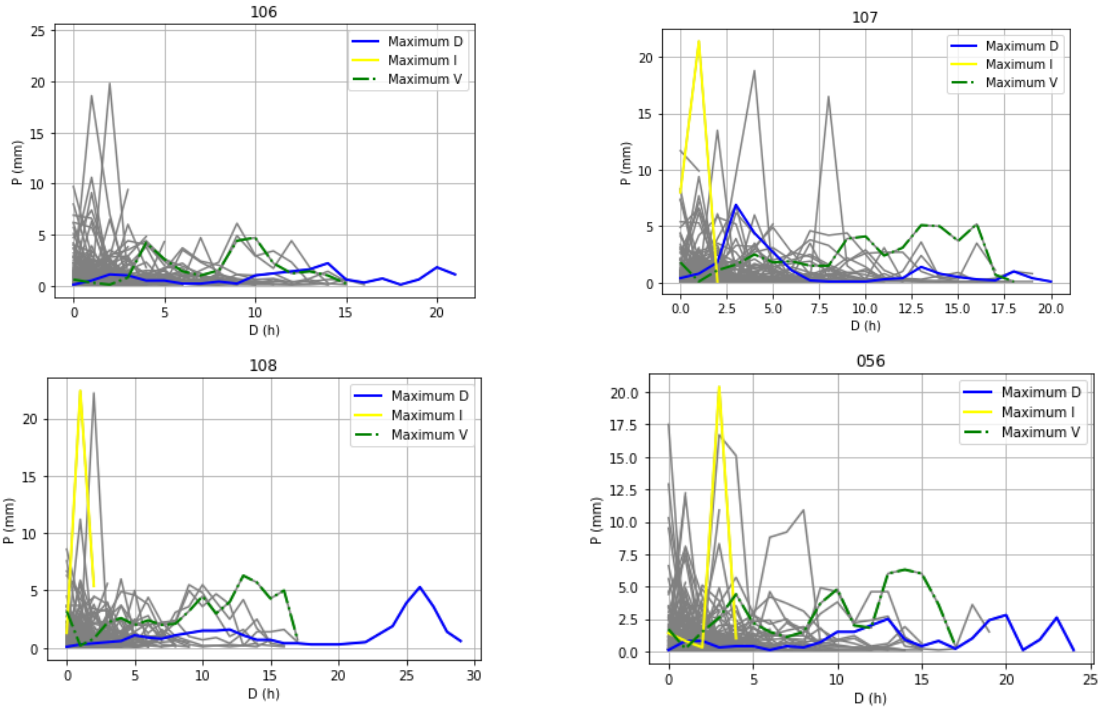
1. Margin	2. Mean	3. Standard deviation	4. Absolute record
5. Recorded storm events (number)	6.	7.	8. 3148
9. Duration (h)	10. 3.6	11. 2.9	12. 31 (Maximum)
13. Volume (mm)	14. 2.2	15. 4.75	16. 54.2 (Maximum)
17. Maximum rainfall Intensity (mm/h)	18. 1.12	19. 2.32	20. 27.4 (Maximum)
21. Runoff volume (mm)	22. 4.49	23. 2.32	24. 30.5 (Maximum)
25. Maximum runoff Intensity (mm/h)	26. 4.08	27. 2.09	28. 28.23 (Maximum)

3148 storm events were recorded in the 6 weather stations. It can be observed from the data presented in table 1, common storms last 3.6 hours falling 2.2 mm with a maximum intensity of 1.12. Using the aforementioned method for estimating the abstractions, 9% of the storms produced runoff.

Figure 2 displays the evolution of the recorded rain volume over the storm duration for each analysed weather station.

Figure 2. Recorded hyetographs by weather station



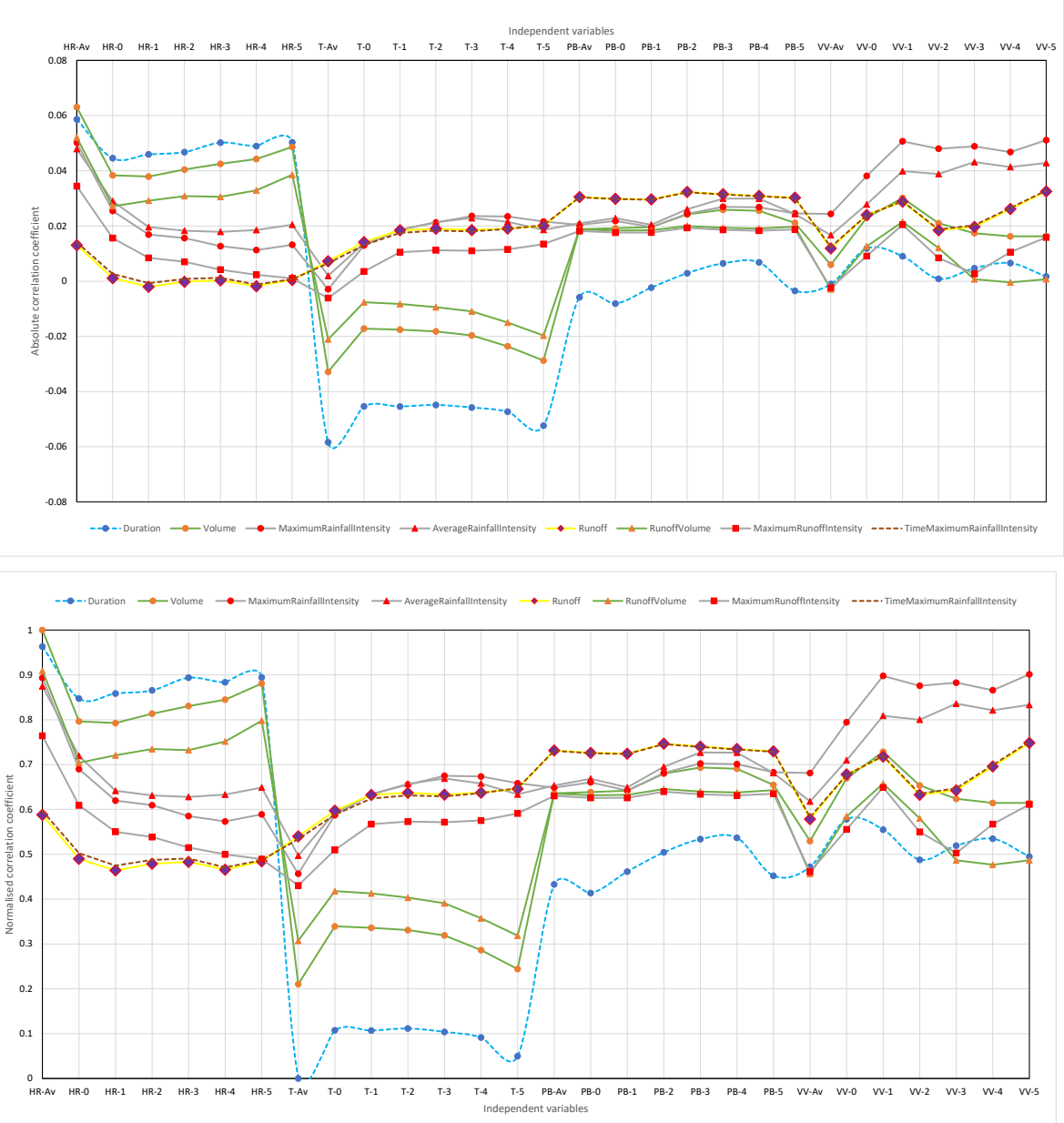


As it can be observed from figure 2, storms producing the maximum rainfall intensity are short while the maximum volumes coincide with long events. Since runoff appears when the rainfall intensity exceeds the maximum infiltration rate of the soil, it can be supposed that the majority of the storm events producing runoff (only 9% of the recorded storms) are short.

For each rainstorm event, we have also recorded the evolution of barometric pressure, relative humidity, temperature and wind velocity during the 5 hours before the storm starts (PB-5, PB-4, PB-3, PB-2, PB-1; HR-5, HR-4, HR-3, HR-2, HR-1; T-5, T-4, T-3, T-2, T-1; VV-5, VV-4, VV-3, VV-2, VV-1), during the first storm hour (PB-0, HR-0, T-0, VV-0) and the average values over the storm duration (PB-Av, HR-Av, T-Av, VV-Av).

In figure 3 we have displayed the linear correlation coefficients, both the absolute and the normalised values, between the potentially dependent and independent variables.

Figure 3. Linear correlation coefficients between dependent and independent variables



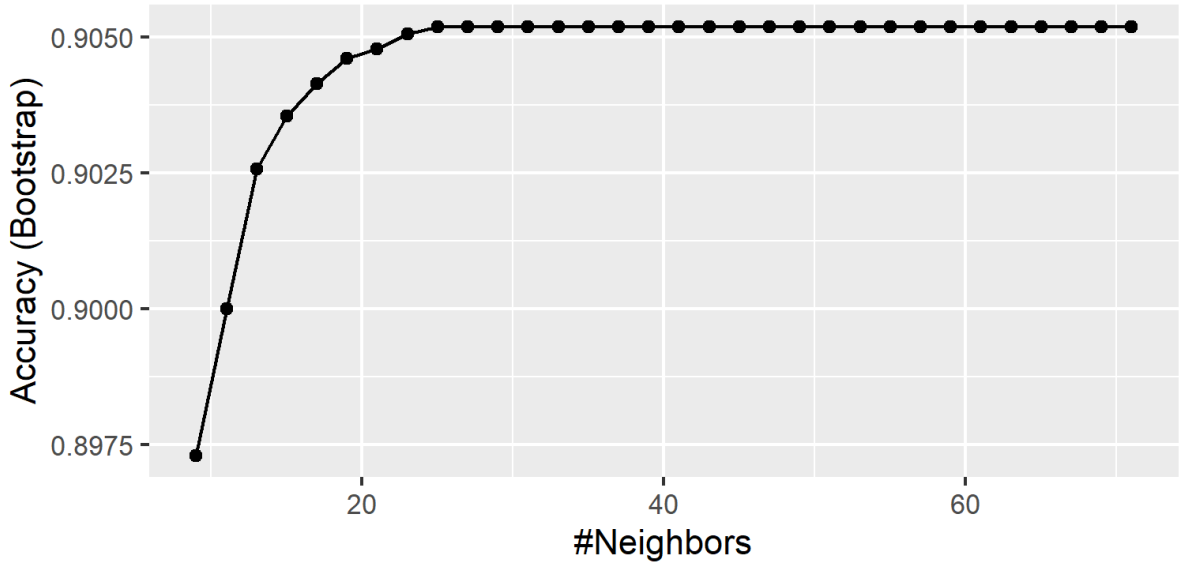
From figure 3 it can be concluded that there are not strong linear relationships between both type of variables and the linear correlation between the runoff appearance and any supposed independent variable is almost zero. The normalized coefficients show that relative humidity, in particular the average relative humidity during the storm event, has the closest linear relationship with volume related variables and wind velocity with intensity related ones. The runoff appearance would be better explained, in linear terms, by the barometric pressure.

From the previous initial exploratory analysis one can think in non-parametric classification algorithms since they are expected to perform better for highly complex non-linear problems as the one faced in this work.

3.2 Classification of storm events with KNN algorithm

Figure 4 presents the results we have achieved from different model´s tuning.

Figure 4. Models accuracy for different k parameter´s values



As it can be observer from figure 4, the model performs best for k>23. However, such model configurations drives to algorithms unable to predict runoff episodes as it can be observed from table 2.

Table 2. Accuracy in predicting runoff episodes

29. K parameter	30. Predicted runoff/Observed runoff
31. 2	32. 11/71
33. 4	34. 6/71
35. 6	36. 5/71
37. 8	38. 2/71
39. 10	40. 0/71
41. 12	42. 0/71
43. 14	44. 0/71
45. 16	46. 0/71
47. 18	48. 0/71
49. 20	50. 0/71
51. 22	52. 0/71
53. 24	54. 0/71

As defined the model does not seem to be valid for predicting the runoff episodes. Some circumstances can explain this fact: there are not enough runoff values to train the model properly (only 9% of the recorded storms produced runoff according to the model used), the

model parameter tuning can be improved by different data pre-processing techniques or by a better strategy for splitting the data between training and testing datasheets.

We achieved similar results when feeding the model only with relative humidity, barometric pressure, temperature or wind velocity individually

5. Conclusions

Merging data-driven algorithms and physically-based models for hydrology looks promising for improving the understanding of hydrological processes. In this work we present the initial results of coupling KNN algorithm and Green-Ampt model for infiltration to estimate the runoff occurrence for a set of recorded storm events.

We have achieved high accuracy when observing the predicting capacity but the models clearly underestimate the runoff occurrence. Further investigation must be conducted to improve the accuracy of the model outputs by retrieving larger datasheets and/or providing the models with more efficient pre-processing or data management strategies.

6. References

Ali, M., Prasad, R., Xiang, Y., & Yaseen, Z.M. (2020) Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *Journal of hydrology*, 584, 124647, https://doi.org/10.1016/j.jhydrol.2020.124647.

Blume, T., van Meerveld, I., & Weiler, M. (2017) The role of experimental work in hydrological sciences-insights from a community survey. *Hydrological Sciences Journal*, 62(3), 334–337.

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., ...& Renner, M. (2017) Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological Sciences Journal*, 64(10), 1141–1158.

Bui, Q.T., Nguyen, Q.H., Nguyen, X.L., Pham, V.D., Nguyen, H.D., & Pham, V.M. (2020) Verification of novel integrations of swarm intelligence algorithms into deep learning neural network for flood susceptibility mapping. *Journal of Hydrology,* 581, 124379, https://doi.org/10.1016/j.jhydrol.2019.124379.

Farajzadeh, J., & Alizadeh, F.A. (2003) hybrid linear–nonlinear approach to predict the monthly rainfall over the Urmia Lake watershed using wavelet-SARIMAX-LSSVM conjugated model. *Journal of Applied Meteorology and Climatology*, 42 381–388.

Fox, N., & Wikle, C. A. (2005) A bayesian quantitative precipitation nowcast scheme. *Weather Forecast*, 20 264–275.

Ivkovic´, M., Todorovic´, A., & Plavšic´, J. (2018) Improved input to distributed hydrologic model in areas with sparse subdaily rainfall data using multivariate daily rainfall disaggregation. *Journal of Hydroinformatics*, 20(4) 784–797.

Mueller, C., Saxen, T., Roberts, R., Wilson, J., Betancourt, T., Dettling, S., Oien, N., Yee, J. (2003) Ncar auto-nowcast system. *Weather Forecast*, 18 545–561.

Niu, W.J., Feng, Z.K., Chen, Y.B., Zhang, H.R., & Cheng, C.T. (2020) Annual streamflow time series prediction using extreme learning machine based on gravitational search algorithm and variational mode decomposition. J. Hydrol. Eng. 2020, 25(5), 04020008, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001902.

Pourghasemi, H.R., Razavi-Termeh, S.V., Kariminejad, N., Hong, H., & Chen, W. F. (2020) An assessment of metaheuristic approaches for flood assessment. *Journal of Hydrology* 582, 124536, https://doi.org/10.1016/j.jhydrol.2019.124536.

Rasmussen, R., Dixon, M., Vasiloff, S., Hage, F., Knight, S., Vivekanandan, J., & Xu, M. (2003) Snow nowcasting using a real-time correlation of radar reflectivity with snow gauge accumulation. *Journal of Applied Meteorology and Climatology*, 42 20–36.

Ruzanski, E., Chandrasekar, V., & Wang, Y. (2011) The casa nowcasting system. *Journal of Atmospheric and Oceanic Technology*, 28 640–655.

Ryu, S., Lyu, G., Do, Y., & Lee, G. (2020) Improved rainfall nowcasting using Burgers' equation. *Journal of Hydrology* 581, 124140, https://doi.org/10.1016/j.jhydrol.2019.124140.

Seed, A. A. (2003) dynamic and spatial scaling approach to advection forecasting. *Journal of Hydroinformatics*, 42 381–388.

Seed, A. Predictability of precipitation from continental radar images. Part iii: operational nowcasting implementation (maple). *Journal of Applied Meteorology and Climatology*, 43 231–248.

Shabri, A., & Suhartono, S. (2012) Streamflow forecasting using least-squares support vector machines. *Hydrological Sciences Journal*, 57(7) 1275–1293.

Tikhamarine, Y., Souag-Gamane, D., Ahmed, A.N., Kisi, O., & El-Shafie, A. (2019) Improving artificial intelligence models accuracy for monthly streamflow forecasting using grey Wolf optimization (GWO) algorithm. *Journal of Hydrology* 582, 124435, https://doi.org/10.1016/j.jhydrol.2019.124435.

Wang, Y., Fang, Z., Hong, H., & Peng, L. (2020) Flood susceptibility mapping using convolutional neural network frameworks. *Journal of Hydrology,* 124482, https://doi.org/10.1016/j.jhydrol.2019.124482.

Zhang, J., Chen, J., Li, X., Chen, H., Xie, P., & Li, W. (2020) Combining postprocessed ensemble weather forecasts and multiple hydrological models for ensemble streamflow predictions. J. Hydrol. Eng. 2020, 25(1), 04019060, https://doi.org/10.1061/(ASCE)HE.1943-5584.0001871.

Zuo, G., Luo, J., Wang, N., Lian, Y., & He, X. (2020) Two-stage variational mode decomposition and support vector regression for streamflow forecasting. *Hydrology and Earth System Sciences*, 24(11), 5491–5518.

Communication aligned with the Sustainable Development Objectives