Machine learning prediction of groundwater heights from passive seismic wavefield

Anthony Abi Nader, J. Albaric, M. Steinmann, C. Hibert, J-P Malet, C. Sue, B. Fores, A. Marchand, M. Gros, H. Celle, et al.

Geophysical Journal International

Geophys. J. Int. (2023) **234,** 1807–1818
Advance Access publication 2023 April 14
GJI Seismology

https://doi.org/10.1093/gji/ggad160

Machine learning prediction of groundwater heights from passive seismic wavefield

A. Abi Nader[1], J. Albaric,[1] M. Steinmann,[1] C. Hibert[2], J.-P. Malet[2,3] C. Sue,[1,4] B. Fores,[1] A. Marchand,[1] M. Gros,[1] H. Celle,[1] B. Pohl,[5] V. Stefani[1] and A. Boetsch[1]

[1] *Chrono-environnement UMR6249, CNRS / Université de Franche-Comté, Besançon, France. E-mail: anthony.abi_nader@univ-fcomte.fr*
[2] *Institut Terre et Environnement de Strasbourg UMR7063, CNRS / Université de Strasbourg, Strasbourg, France*
[3] *Ecole et Observatoire des Sciences de la Terre UAR830, CNRS / Université de Strasbourg, Strasbourg, France*
[4] *Institut des Sciences de la Terre (ISTerre), Université Grenoble Alpes, Université Savoie Mont Blanc, CNRS, IRD, Université Gustave Eiffel, Grenoble, France*
[5] *Biogéosciences UMR6282, CNRS / Université de Bourgogne, Dijon, France*

Accepted 2023 April 12. Received 2023 February 7; in original form 2022 October 3

SUMMARY

Most of water reservoirs are underground and therefore challenging to monitor. This is particularly the case of karst aquifers which knowledge is mostly based on sparse spatial and temporal observations. In this study, we propose a new approach, based on a supervised machine learning algorithm, the Random Forests, and continuous seismic noise records, that allows the prediction of the underground river water height. The study site is a karst aquifer in the Jura Mountains (France). An underground river is accessible through an artificial shaft and is instrumented by a hydrological probe. The seismic noise generated by the river is recorded by two broadband seismometers, located underground (20 m depth) and at the surface. The algorithm succeeds in predicting water height thanks to signal energy features. Even weak river-induced noise such as recorded at the surface can be detected and used by the algorithm. Its efficiency, expressed by the Nash–Sutcliffe criterion, is above 95 per cent and 53 per cent for data from the underground and surface seismic stations, respectively.

Key words: Hydrogeophysics; Seismic noise; Machine learning; Time-series analysis.

1 INTRODUCTION

Water resource has become an essential environmental and societal issue due to the intensification of its exploitation and its vulnerability to climate change (Drew 1999; Andreo *et al.* 2006; Green *et al.* 2011). Drinking water supply relies mainly on groundwater aquifers, which are generally not directly discernible nor accessible (Chen *et al.* 2017; McDonnell 2017). This applies in particular to karst aquifers, which are very heterogeneous in terms of permeability : they are characterized by fast groundwater flows in open conduits (underground rivers) and slow flows in the micro-fractured rock matrix (Ford & Williams 2013). Since most karst aquifers are inaccessible, their monitoring often relies on punctual observations from piezometers or on spring hydrographs. In order to better understand these systems, it is therefore essential to develop new monitoring approaches, adapted to their heterogeneous geometry and flow dynamics.

The seismic wavefield has proved to provide information about hydrogeological processes (e.g., Larose *et al.* 2015). Actually, the hydrodynamics of surface rivers have been the subject of several passive seismic studies. For example, the spectral analysis of the ambient seismic noise induced by river flow has allowed to identify sediment transport and deposition within stream segments (Burtin *et al.* 2008; Schmandt *et al.* 2013). It has been shown that seismic recordings from geophones installed on the river bank could be used to estimate river discharge (Anthony *et al.* 2018), water height and bedload transport (Dietze *et al.* 2019). Regarding groundwater, seismic interferometry methods are effective in detecting water level within the rock matrix, through the measurement of seismic velocity changes in the subsurface (Voisin *et al.* 2017; Fores *et al.* 2018; Vidal *et al.* 2021). Measuring hydrogeological parameters of underground river, which are generally inaccessible, remains however challenging.

Recent advances in research combining machine learning and seismic monitoring have shown that it is possible to identify automatically the sources of seismological events triggered by various geological processes. Actually, the Random Forest algorithm and curated features have been successful in describing landslide microseismicity (Provost *et al.* 2017; Wenner *et al.* 2021), differentiating between rockfalls and volcano-tectonic earthquakes (Hibert *et al.* 2017), detecting debris flow events (Chmiel *et al.* 2021) and establishing seismic lithofacies classification (Kim *et al.* 2018). The Random Forest algorithm can also be used to predict, in the machine learning term, continuous values. For example the method was

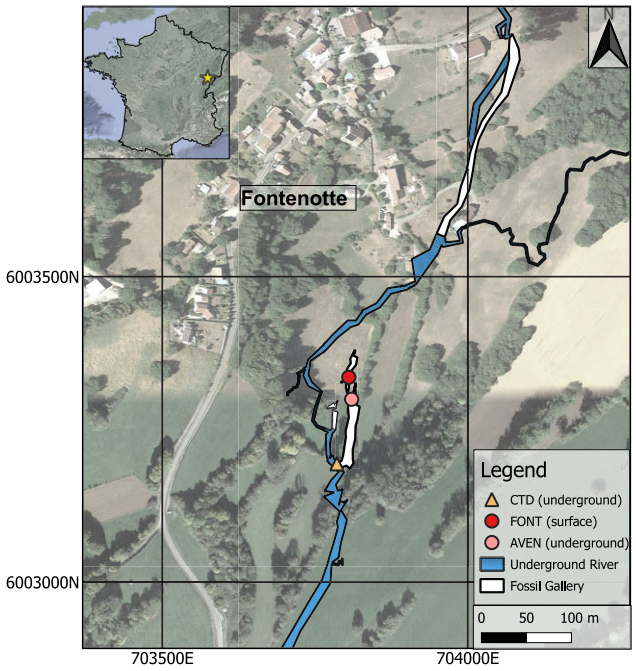

Figure 1. Location map of the site and instrumentation. The underground river and the dry fossil galleries are drawn in blue and white, respectively (topographical data are from D. Motte, ASDC). The red and pink dots show the positions of the seismic stations and the triangle the position of the hydrogeological probe (CTD).

applied on laboratory observations to identify hidden signals that precede earthquakes and predict the time remaining before failure (Rouet-Leduc *et al.* 2017). It was also applied to predict subsurface porosity and capture its spatial variation in reservoirs based on seismic attributes (Zou *et al.* 2021).

The objective of this study is to propose an innovative approach based on a continuous application of the Random Forest machine learning algorithm on passive seismic wavefield to provide a remote inference of the water height of the underground river. In other terms, we propose to establish a projection of seismic data in a multi-dimensional feature space extracted using a 15-min-long sliding window to the output of 1-D water height values using the mentioned algorithm. This unprecedented application could be the head start in the investigation of other inaccessible water conduits towards a better groundwater estimation and flood forecasting.

2 STUDY SITE AND DATA

The study site is the Fourbanne karst aquifer in the Jura mountains, eastern France (Fig. 1). It is part of the JURASSIC KARST hydrogeological observatory settled in 2014 (Cholet *et al.* 2017) and of the french SNO KARST network (Jourde *et al.* 2018). The local lithology is characterized by Middle Jurassic tabular limestones and shales cross-cut by a series of N-S and NE-SW normal faults, which control the orientation of the underground conduits. The aquifer is primarily fed by allogenic recharge through sinkholes (Cholet *et al.* 2015). The underground conduit has been explored and mapped over a length of 9 km by speleologists in the unsaturated zone and by cave divers in the saturated zone. The location of the instruments as well as a part of the karst conduit are detailed in Fig. 1. The seismological data are recorded by two stations of the long-term regional seismic network JURAQUAKE deployed in eastern France since late 2018 and 2019. The station AVEN is located in a fossil gallery at 20 m depth (423 m asl), at the base of a vertical shaft drilled by speleologists (Guralp CMG40T 60s-100Hz sensor, connected to a Staneo D3BB-MOB digitizer). The second station FONT is located at the surface (443 m asl), at 3 m from the well-head (Guralp 6TD, 30s-100Hz sensor). For coupling purposes both seismometers are dug 50 cm into the cave sediment or surface soil. AVEN and FONT are at a slope distance of about 50 m and 60 m from the underground river's channel. The sampling frequency of these three component seismic stations is 200 Hz for AVEN and 100 Hz for FONT. A hydrological probe (CTD) is installed in the river and records water electrical conductivity, water temperature and water height every 5 minutes.

In this study, we focus on hydrogeological data recorded for 2 years between 2009 September 15 and 2021 September 15 (Fig. 2). This period covers two entire hydrogeological cycles, with main rainy seasons in winter and spring. During this period, the CTD recorded a minimal water height of 0.4 m during low water periods, which is measured from the streambed to the water–air interface, and a maximum height of 1.7 m during floods. Seismological data are complete during this period of time at AVEN only. Due to technical problems, there are gaps in the data recorded at FONT and the analysis covers a shorter period of time: between 2019 October 27 and December 31 and between 2020 September 15 and 2021 September 15. Fig. 2(a) is a plot of the underground river water height during all of the studied period. Fig. 2(b) is a zoom on a flood occurring between 2019 November 16 and November 20. Spectrograms computed from seismological data recorded at FONT and AVEN, during this same flood, are presented in Figs 2(c) and (d), respectively. Energy lines between 10 and 20 Hz appearing on both AVEN's and FONT's spectrograms can be related to the anthropogenic activity. Indeed, these energy lines are more marked for FONT than for AVEN due to its location underground, insulated from the surface, thus the river induced noise will appear on its signals' spectrogram with a higher amplitude. Actually, we can notice at this frequency range a day-night variation with more energy during daytime, and less energy during days-off (November 17 is a Sunday). In addition, for the latter frequency range, more energy is manifested on the horizontal components than on the vertical component (Fig. A1).

The seismic noise induced by the river becomes relatively visible on the spectrograms once the water rises. Due to the position of the seismometer, at the surface, in a field enclosing two horses, and the prominence of noise generated by anthropogenic sources (vehicles on the nearby road, dwellings, agricultural activities, mining) as well as the horses' gaits, the effect of water height variation is hardly detectable on the FONT spectrogram (Fig. 2c). The noise amplitude increase due to water height increase is clearer on the AVEN spectrogram, which is more isolated from the surface noise (Fig. 2d). Actually, three main frequency ranges can be associated with water height change (Fig. 2d): 1–3 Hz, 5–8 Hz and 25–50 Hz. While the low frequency bands (1–8 Hz) are visible before the flood, energy at high frequency seems to occur after the flood has started. The seismic noise related to water flow in rivers can be associated to different phenomena (Burtin *et al.* 2008; Tsai *et al.* 2012; Schmandt *et al.* 2013; Díaz *et al.* 2014; Gimbert *et al.* 2014). A major source of noise results from the frictional forces produced by the interaction between the turbulent flow and the riverbed. Another one is the bed load particles transport, generally observed at higher frequency.

In order to look more in detail at the noise induced by the flood, we have plotted the noise amplitude against water height with a 5 min time step for FONT (Figs 3a–c) and AVEN (Figs 3d–f). Data were filtered at 1–3 Hz, 5–8 Hz and 25–50 Hz, corresponding to the

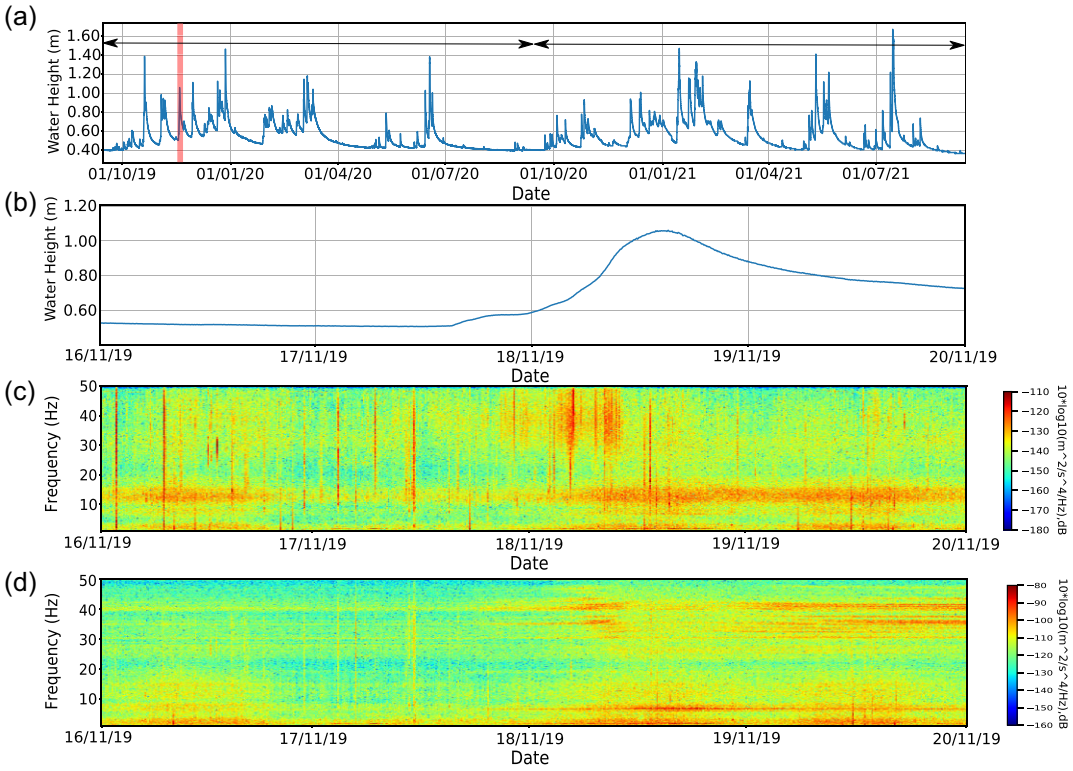

Figure 2. (a) Hydrograph spanning the entire study period in terms of water height. The shaded part is the flood selected for the spectrograms computation. The two arrows indicate the two hydrogeological cycles used for the training and application of the RF algorithm. (b) Hydrograph of the selected flood event in (a), between 2019 November 16 and 20. (c) Spectrogram of the vertical components during the selected flood for signals recorded at FONT and filtered between 1 and 50 Hz. (d) Same as (c) for AVEN.

three frequency ranges evoked earlier. Similar trends are obtained for the two stations. While the water height rises from 0.50 m to 0.80 m and decreases below 0.80 m, the variation of the noise versus the variation of water height follow the same path. In addition, at all frequency ranges, between 0.90 m and the flood peak, the noise amplitude as function of water height draws a hysteresis. The latter is generally attributed to bedload transport during increasing flooding (high noise amplitude), and gravel deposit during recession (lower noise amplitude), with lower frequencies corresponding to larger particle movement (Burtin *et al.* 2008; Schmandt *et al.* 2013; Díaz *et al.* 2014). Gimbert *et al.* (2014) have also shown that turbulence processes in the river could also significantly contribute to the hysteresis curve.

3 FEATURE EXTRACTION FROM SEISMIC DATA

In this section the objective is to extract characteristic features from the raw seismic data. Firstly, a pre-processing of the seismic data has been performed. The data are decimated to 100 Hz (for AVEN only), and then detrended and filtered between 1 and 50 Hz. The seismic signals are then partitioned using a 15 minutes moving window with an overlap of 50 per cent, corresponding to a windowing step of 7.5 min. Several window lengths were tested. A 15-min-long window was chosen because it requires reasonable CPU time computation and provides sufficient resolution for capturing the beginning of the rise of water during a flood event. Finally, features of the seismic recordings are computed for each window (Table 1, see Hibert *et al.* 2017, for a detailed description of each feature). A total of 72 features are calculated related to the signals' waveform, frequency content, spectral energy, and pseudo-spectrogram. Similar features as Hibert *et al.* (2017) are used in our study, except for the polarity attributes, with additional frequency bands for the computation of the signal's Kurtosis and energy (1–3, 3–5, 5–8, 8–10, 10–15, 15–20, 20–25, 25–30, 30–35, 35–40, 40–45, 45–50 Hz). These frequency bands are chosen to cover all of the studied frequency range (1–50 Hz) and target the bands affected by the water height variation obtained during the spectral analysis. These features, as presented in Hibert *et al.* (2017), are commonly used to identify events or seismic sources within the seismic signals since they are able to cover several aspects of the signals. In the case of the chosen configuration, the computation of features for a station and a year of data takes about 2 weeks of CPU time. The extracted features are used in the algorithm that is explained in the following section.

4 METHOD

The Random Forest (RF) algorithm (Breiman 2001) is a bagging ensemble learning method based on the computation of a large number of decision trees. Each tree in the forest is generated from a random subset of events from the training set and a random subset of features describing the events. The RF algorithm has two modes of application: (1) the classification in which the final result will be a class obtained from the majority of voting, and (2) the regression in which the final result will be a value obtained by averaging the predicted values given by each tree. Increasing the number of trees in the forest helps in the convergence without causing overfitting but reducing the generalization error (Breiman 2001), which

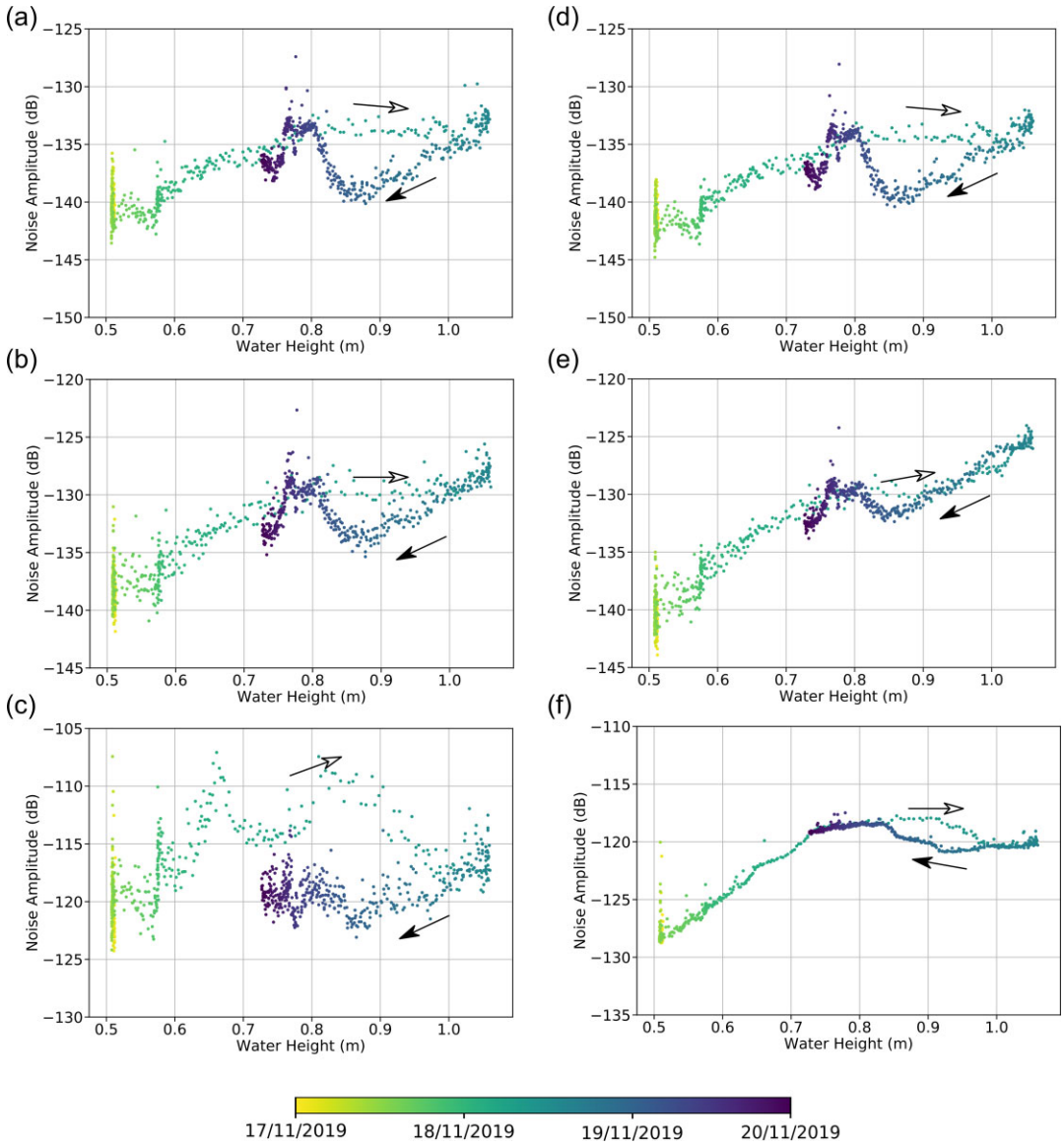

Figure 3. Noise amplitude as function of the water height for signals recorded at the surface station FONT and filtered between (a) 1–3 Hz, (b) 5–8 HZ and (c) 25–50 Hz. Same for underground station AVEN in (d–f). The color bar represents the time scale between 2019 November 17 and November 20. The white and black arrows correspond respectively to the increase and decrease sections of the water height during the flood.

measures the prediction error of the model over the data set. Because of the random selection of both the training events and features, each decision tree in the Random Forest is unique and the trees are not correlated with each other that helps in reducing the overfitting, which is one of the advantages of this algorithm. Another advantage is the ability of the algorithm to use a large number of features and assess their importance depending on the attributed case (water height in this study) in the prediction, while going from a multidimensional dataset to a 1D output. The importance of the features helps to better understand the results and provide insights on the link between the seismic signals features and the physics of the phenomena.

Each decision tree of the RF consists of internal nodes (splits) and terminal nodes (leaves) (Criminisi *et al.* 2011). The depth of a tree is the number of splits from its root (node 0) to its leaves. It is a measure of the number of splits made by the tree to get a prediction. No limitation was set on the maximum depth in our model: the nodes are expanded until all leaves contain less than two samples in the population during the splitting. The deeper the tree, the more splits it has, hence more information will be caught from the data and configured into the model. At each node, the selected feature is used to split the selected subset of data into two separate populations. The best splitting value at each node is found by variance reduction, meaning the value of the feature at the split is the value giving the lowest variance between the predicted (which is the mean value of each obtained population) and the real values, and thus yielding the highest precision. Feature importance is the assignment of a score to features based on their impact on the targeted prediction. The feature giving the lowest variance in the splitting is the feature with the highest importance and is chosen as a root node for the tree. The objective of the feature importance is to assess the influence of each feature on the model decision making in order to interpret the resulted predictions. Another objective can be to select important features for similar applications and gain in processing time.

Table 1. Description of the features used in the algorithm (modified from Hibert *et al.* 2017). DFT and FFT stand for Discrete Fourier Transform and Fast Fourier Transform respectively.

Feature	Description
Duration	Duration of the signal
RappMaxMean	Ratio of the Max to the mean of the normalized envelope
RappMaxMedian	Ratio of the Max to the median of the normalized envelope
AsDec	Ratio of the ascending to decreasing time of the envelope
KurtoSig	Kurtosis of the signal
KurtoEnv	Kurtosis of the envelope
SkewnessSig	Skewness of the signal
SkewnessEnv	Skewness of the envelope
CorPeakNumber	Number of peaks in the autocorrelation function
INT1	Energy in the first 1/3 of the autocorrelation function
INT2	Energy in the last 2/3 of the autocorrelation function
INT_RATIO	Ratio of INT1 to INT2
ESi-j	Energy of the seismic signal in the i-j Hz frequency band
Kurtoi-j	Kurtosis of the signal in the i-j Hz frequency band
DistDecAmpEnv	Difference between decreasing coda amplitude and straight line
RatioEnvDur	Ratio between maximum envelope and duration
MeanFFT	Mean FFT
MaxFFT	Max FFT
FmaxFFT	Frequency at Max (FFT)
FCentroid	Frequency of spectrum centroid
Fquart1	Frequency of 1st quartile
Fquart3	Frequency of 3rd quartile
MedianFFT	Median of the normalized FFT spectrum
VarFFT	Variance of the normalized FFT spectrum
NpeakFFT	Number of peaks in the normalized FFT spectrum
MeanPeaksFFT	Mean peaks value for peaks >0.7
E1FFT	Energy in the 1 – NyF/4 Hz (NyF = Nyqusit Frequency) band
E2FFT	Energy in the NyF/4 – NyF/2 Hz band
E3FFT	Energy in the NyF/2–3∗NyF/4 Hz band
E4FFT	Energy in the 3∗NyF/4 – NyF/2 Hz band
gamma1	Spectrum centroid
gamma2	Spectrum gyration radio
gammas	Spectrum centroid width
SpecKurtoMaxEnv	Kurtosis of the envelope of the maximum energy of spectrograms
SpecKurtoMedianEnv	Kurtosis of the envelope of the median energy of spectrograms
Ratioenvspecmaxmean	Ratio of the Max DFT(t) to the mean DFT(t)
Ratioenvspecmaxmedian	Ratio of the Max DFT(t) to the median DFT(t)
Distmaxmean	Mean distance bewteen Max DFT(t) mean DFT(t)
Distmaxmedian	Mean distance bewteen Max DFT median DFT
Nbrpeakmax	Number of peaks in Max (DFTs(t))
Nbrpeakmean	Number of peaks in mean (DFTs(t))
Nbrpeakmedian	Number of peaks in median (DFTs(t))
Rationbrpeakmaxmean	Ratio between the number of peaks in Max (DFTs(t)) and mean (DFTs(t))
Rationbrpeakmaxmed	Ratio between the number of peaks in Max (DFTs(t)) and Median (DFTs(t))
Nbrpeakfreqcenter	Number of peaks in centroid frequency DFTs(t)
Nbrpeakfreqmax	Number of peaks in Max frequency DFTs(t)
Rationbrfreqpeaks	Ratio between the number of peaks in centroid frequency DFTs(t) and Max frequency DFTs(t)
DISTQ2Q1	Distance Q2 curve to Q1 curve (QX curve = envelope of X quartile of DTFs)
DISTQ3Q2	Distance Q3 curve to Q2 curve
DISTQ3Q1	Distance Q3 curve to Q1 curve

An example of a tree of the forest resulting from the training of a model with AVEN's data with the maximum depth set at 3 is presented in Fig. 4. This maximum depth is only used to generate this figure and to be able to visualize the functioning of a tree. A subset of features and data are selected for this tree. The features are sorted according to their importance. At the level of each node, water height is plotted as a function of the feature corresponding to the node. The feature corresponding to the energy of the seismic signal between 40 and 45 Hz (ES40–45), which represents the base 10 logarithm of the integral of the raw seismic signal's envelope filtered between 40 and 45 Hz, is the root of the tree since it is the most important feature. A splitting point for this feature is obtained and the population is divided into two sets accordingly, each set having its own mean water height that gives the lowest variance. Samples having this feature above the resulted threshold will be selected at the right part of the tree, and below the threshold at the left part. The next feature splitting the population will be less important than the preceding feature. This is done at every node until the maximum depth condition is fulfilled. The samples obtained at each node will be satisfying all the above conditions from all the previous nodes of the tree.The final plots represent the water height distribution at different time windows of the remaining samples after

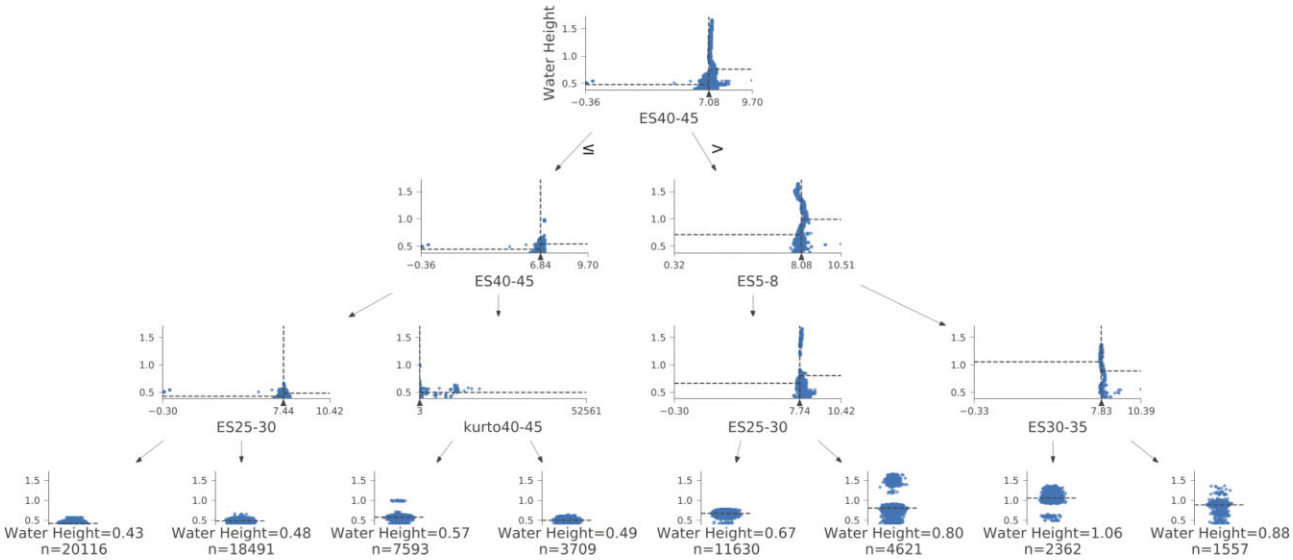

Figure 4. A tree diagram generated from a model using AVEN's data with a maximum depth parameter set at 3.

the splitting. At the level of each final node is the number of samples obtained for the final split and the prediction of the water height, which is the average value of the water heights of the remaining samples. Given features corresponding to a certain signal, a water height will be obtained at each tree, and having 1000 trees the final prediction will be the mean of all obtained water heights.

Unlike usual applications, the training and testing are here done on continuous signals (water height and seismic) and not on selected events. The process involves independent algorithms for each station (AVEN and FONT). The following steps include training on a certain period of time and testing on another period of time. The choice of these periods of time was controlled by data availability. The training period is the same for AVEN and FONT: from 202 September 15 to 2021 September 15 (training dataset). The testing period (testing dataset) is from 2019 September 15 to 2020 September 15 AVEN and only two months for FONT (from 2019 October 27 to December 31). Hence the training dataset counts 70 000 windows (of 15 min) for AVEN and FONT, and the testing dataset counts 70 000 windows for AVEN and 13 000 windows for FONT. The choice of training the algorithm on 2020–2021 data and testing it on 2019–2020 instead of the other way round is the lack of data for FONT; this choice allows to have firstly the data of a complete hydrological cycle for the training to cover all potential water heights and secondly a same training period for both stations for results comparison issues. A RF with 1000 trees is then created based on the training dataset by assigning to each window of features the corresponding measured water height. A similar configuration was used as in Provost *et al.* (2017) and Hibert *et al.* (2017). The RF model was then applied on the testing dataset which generates an array of water heights. The predicted water heights are then compared with the real values in order to assess the algorithm precision. Finally, to evaluate which features are the most relevant, 10 forests were created and trained, each giving values for the features importance. These values are then averaged over the 10 instances. This number of instances was chosen since it is a reasonable choice in terms of CPU time, knowing that one instance of the algorithm can take several hours since we are using a year of data for the training of a 1000 trees forest.

5 RESULTS OF THE REGRESSION

The results of the regression analysis performed on data from stations FONT and AVEN are presented in Fig. 5. The predicted water height was smoothed for both stations using a 10-day moving window to avoid short transitory signals coming from local noise sources. A good fit between observed and predicted values was obtained, as illustrated in Fig. 5. In order to better assess the quality of the fit we calculated the overall Root Mean Square Error (RMSE) and the Nash–Sutcliffe efficiency coefficient (NSE) which is commonly used in hydrological models (McCuen *et al.* 2006) and is given by:

$$NSE = 1 - \frac{\sum_{t=1}^{T}(H_0^t - H_m^t)^2}{\sum_{t=1}^{T}(H_0^t - \bar{H}_0)^2} \quad (1)$$

where $\bar{H}_0$ is the mean of observed water heights, and H_m is modeled water height. H_0^t is observed water height at time t. For FONT, the obtained RMSE is about 0.1 m and the NSE is about 53 per cent. The prediction shows many outliers or misfits compared to the true water height variation. The misfits are mainly observed during the recession period (e.g. beginning of December 2019; Figs 5a and c). It is most likely due to the position of the station at the surface, at the vicinity of many major noise sources which tends to hide the noise generated by the river. However, the overall shape of the hydrograph is correctly reproduced. For AVEN the quality of the fit is very high, with an RMSE of only 0.03 m and the NSE reaching 95 per cent. A few outsider peaks can be observed systematically during periods of flood recession : mostly from mid to end of 209 January, from 2019 mid-March to end of April and from beginning of 2019 July to end of September. For these periods the predicted water heights fall mainly below and only punctually above the observed heights (Figs 5b and d). In these cases, the seismic noise generated by the river most likely interferes with other noise sources. Figs 5(c) and (d) show the predicted values of the water height versus the real measured values on a normalized scale for simulations done with FONT and AVEN, respectively. If the fit between predicted and

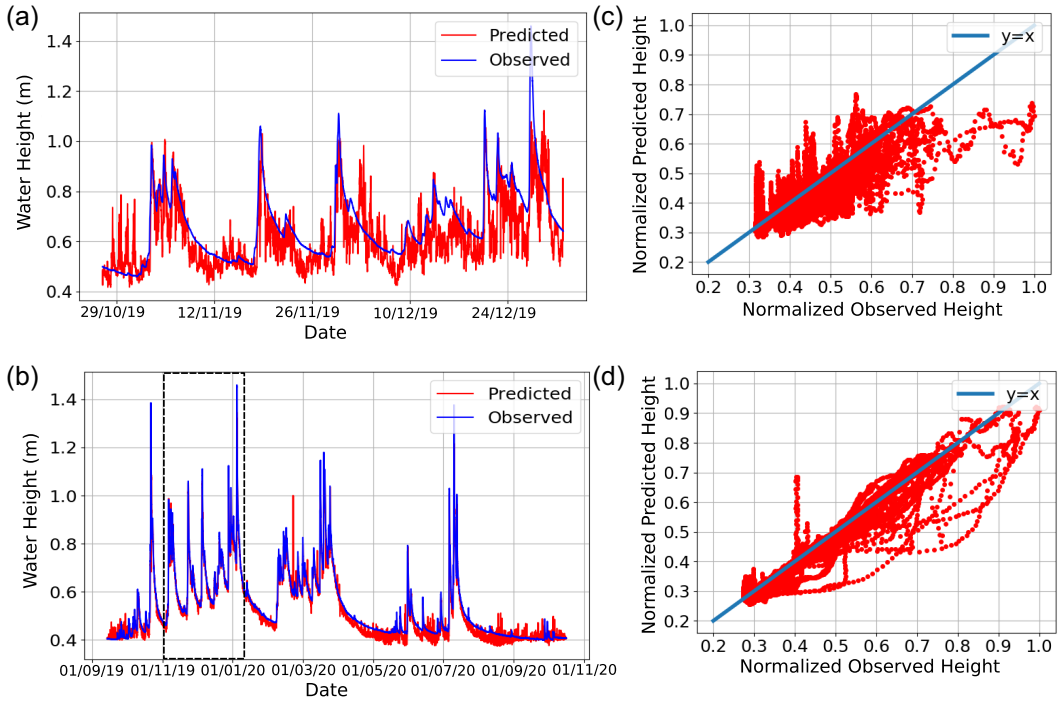

Figure 5. (a) Simulated water height at surface station FONT; the blue line is the water height measured at the CTD; the red line is the predicted water height obtained from the application of the algorithm on the seismological data. (b) Same as (a) for underground station AVEN. The dashed rectangle indicates the period of application in (a). (c) Predicted versus measured water height for simulations carried out at FONT; the blue line represents the 1:1 line. (d) Same as (c) for AVEN.

real values was perfect they would have been aligned along the 1:1 line. We observe that for both stations, most of the points are below the 1:1 line, meaning that the predicted water height is lower than the observed one. This could be due to the difficulty to predict the water heights during low water periods, when the induced noise amplitude is lower. We also computed the Pearson coefficient (R) for both applications to measure the strength and direction of the linear relationship between the predicted and observed values. As expected, the coefficient is high for AVEN (0.98) and lower for FONT (0.70).

6 DISCUSSION

The features importance resulting from averaging scores over 10 instances of training are presented in Fig. 6. The most important features retained by the algorithm are waveform and signal energy features for both stations. In the case of FONT, the dominant feature is the Kurtosis between 8 and 20 Hz (a plot showing the variation of the kurtosis feature for signals filtered between 15 and 20 Hz compared with the water height is shown in Fig. B2 of the appendix). In the case of AVEN, signal energy is the most important feature, for frequencies between 5 and 8 Hz and at high frequency (above 20 Hz), especially between 40 and 45 Hz (a plot showing the variation of the signal's energy feature between 40 and 45 Hz compared with the water height is shown in Fig. B1 of the appendix).

The Kurtosis of the seismic signal is a common feature used in many classification applications which allows detecting natural or anthropogenic seismic events within continuous seismic records (e.g. Liang *et al.* 2008; Baillard *et al.* 2013; Ross & Ben-Zion 2014). This may justify why this feature is the most important for the FONT regression model. The position of this station at the surface near roads and agricultural fields leads to the presence of random punctual peaks in the seismic signals thus causing an increase in the kurtosis. Since we are using 15 min windows, multiple few-seconds events can be included leading to a flatter normal distribution and a higher kurtosis value, which can be detrimental to the variation of the kurtosis with the water height. The skewness of the seismic signal is also a common feature used for event detection, especially seismic phase detection (Ma *et al.* 2015; Küperkoch *et al.* 2010). Hence, it might also be affected in our case by anthropogenic events. We tested the algorithm without the kurtosis and the skewness features, which allowed improving the results (Fig. 7).

The predicted hydrograph is smoother, presenting less misfits and similar RMSE (0.14 m) compared with the initial prediction. The NSE is however smaller (13 per cent against 53 per cent). Indeed, the amplitude of the floods are not completely recovered. As mentioned above, the Kurtosis feature is usually used for event detection applications, hence its importance in the detection of floods as well as the detection of transitory water height peaks during a flood event. For example for the flood occuring between the 2019 December 20 and the 26 (Fig. 7), we can notice that the first peak is correctly reproduced, unlike the following peaks, which are underestimated. Applying the algorithm while removing the kurtosis features at FONT allows to highlight the signal's energy feature, in particular between 1–8 and 20–30 Hz (Fig. 8).

According to results at the two stations, the main feature of the seismic noise use by the RF algorithm for simulating the water height is the signal energy. At both stations, these features exhibit

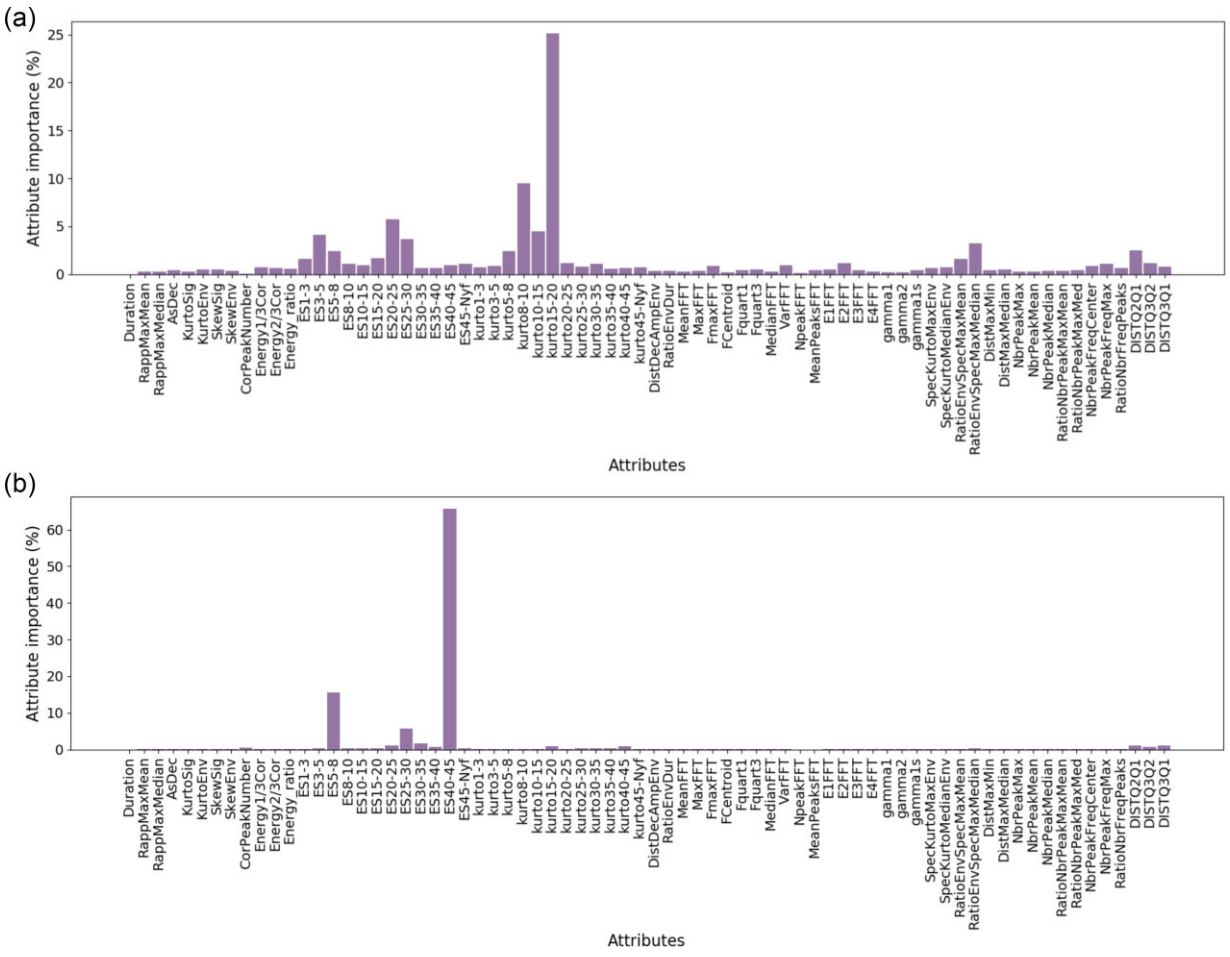

Figure 6. Feature importance obtained by averaging scores over 10 instances of training and testing the RF algorithm for stations (a) FONT and (b) AVEN.

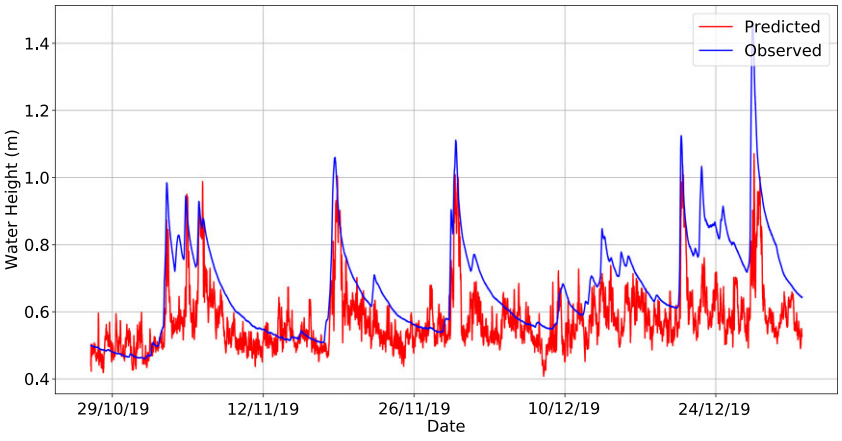

Figure 7. Water heights simulated with FONT's data after removing the kurtosis and skewness features; the blue and red lines correspond to the observed and predicted water height, respectively.

low and high frequency contents, at 1–8 and >20 Hz, respectively. As mentioned in section 2, hydrodynamic processes have a particular spectral signature, which has been described and modeled for surface river (Burtin *et al.* 2008; Tsai *et al.* 2012; Gimbert *et al.* 2014). Observed features are most likely related to these mechanisms : the low frequency band features would correspond to the turbulence, while higher frequency content (>20 Hz) could be due to bed load transport.

Associated frequency bands are more widely distributed at FONT than at AVEN, which could be explained by their respective location (surface and 20 m depth) affecting their ability to capture river induced signals only. In addition, due to seismic attenuation, a station

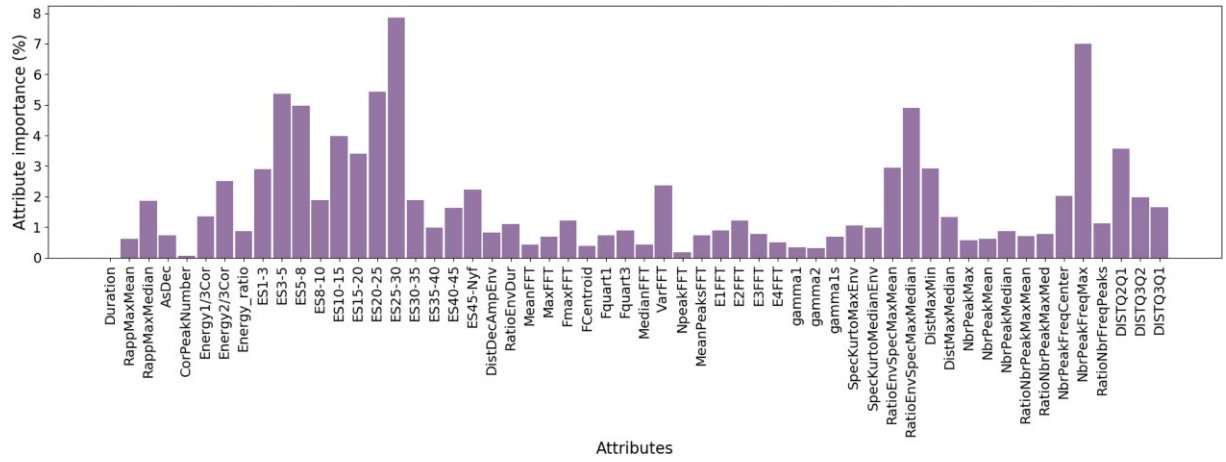

Figure 8. Feature importance obtained by averaging scores over 10 instances of training and testing the RF algorithm for FONT without the kurtosis and skewness features.

closer to the underground river, such as AVEN, is able to record higher frequency content, compared to a surface station such as FONT. The channel morphology at Fontenotte is very heterogeneous (the river width can vary in terms of several meters) which makes physical models developed for surface river (Tsai *et al.* 2012; Gimbert *et al.* 2014) difficult to apply. However, bed load transport model developed by Tsai *et al.* (2012) and field observations provided by Burtin *et al.* (2008) indicate that the smaller is the distance station-river, the higher is the peak frequency. Our results corroborate this, with the second best feature being the seismic energy value in the 25–30 Hz frequency band for the farthest station (FONT - approximate distance of 60 m) and the best feature being the seismic energy fluctuation in the 40–45 Hz for the closest and deepest station (AVEN – approximate distance of 50 m). This highlights the strong correlation between the seismic energy in those frequency bands and the dynamics of the river, and hence explains the ability of the RF machine learning model to reproduce the hydrograph from seismic signals.

One limit of the methodology presented here is the necessity of having hydrogeological data for the training. On the other hand, once the training is complete, no access is anymore needed to the underground river to do the maintenance of the hydrological probe and collect its data. In addition, the hydrological probe should be installed a period long enough to cover all potential water heights that might occur. Indeed, the method is not able to predict water heights that it was not trained for, in other terms it cannot extrapolate. Further investigation to unravel this could be to use data at the river's outlet, where the access is generally easy, in order to train the RF and then test it at other positions on the conduit.

Linear regression is a simple and efficient mean allowing to simulate water height as demonstrated by Anthony *et al.* (2018), who used data from geophones located at the vicinity of the river (1 m from the stream). In this station configuration, the seismic energy is dominated by the noise induced by the river. This is not the case when monitoring an underground river from surface, as observed at FONT where the river-induced signal is most likely too weak and noisy to apply regression analysis. On the other hand, machine learning method can detect the most relevant frequency ranges for performing a successful simulation, even for distant and noisy stations. In fact, multiple features were revealed relevant for the regression, which shows the importance of this method in it's ability to deal with such datasets.

Although the predicted hydrograph fits more correctly the observed values using data from a very high signal to noise ratio sensor closer to the source (such as AVEN), the application of the method using data from the surface station is promising: the prediction fits well the temporal evolution of the real water height permitting the detection of floods as well as roughly estimating the corresponding water height. Better regression for FONT can potentially be obtained by improving the station insulation (e.g. FONT is only buried at 0.5 m depth) which could enhance the signal-to-noise ratio of the recorded signals. Further investigation could be performed to better characterize the effect of the geometry and the dynamics of the river. An approach could be to test the algorithm with data from a denser seismological array installed all along the river as well as at the surface. Finally, the proposed method could be extrapolated to other accessible or inaccessible sites and implemented for real-time applications allowing a continuous monitoring of flood events in order to circumvent the installation of invasive instruments.

7 CONCLUSIONS

We used a Random Forest (RF) algorithm to predict the flow dynamics of an underground karstic river using seismological data. Seventy-two features characterising the seismic signals were computed to train the model and predict the water heights. With an RMSE of 0.03 m obtained for the regression using the data collected from the underground seismological station (AVEN) and of 0.1 m using the data collected from the seismological station at the surface (FONT), the RF model proved to be a reliable method for remote monitoring of the water heights. Feature importance was computed in terms of variance between predicted and observed values. The most important features correspond to the signal energy for AVEN and FONT. They are related to the contribution of the water-riverbed interaction and the bed load transport on the seismic noise content. The results demonstrate the accuracy of the method in predicting underground river water heights, even with weak river-induced and noisy signals (the case for FONT).

ACKNOWLEDGMENTS

This project has been funded by the Région Bourgogne-Franche-Comté and OSU THETA. Calculations were performed using HPC

resources from DNUM CCUB (Centre de Calcul de l'Université de Bourgogne). We are very thankful to the members of the Association Spéléologique du Doubs Central (ASDC) and J.-P. Simonnet (Chrono-environnement, UFC) for their active support in the field work. We thank D. Motte for his scientific contribution and for giving us access to the topographical data of the Fontenotte cavity. We are also thankful to the land owners and tenants of the instrumented site as well as the municipality actors for their help and availability. We also acknowledge the French Karst National Observatory Service (SNO KARST, https://sokarst.org/) and the OZCAR (Critical Zone Observatories: Research and Application) Research Infrastructure. We are grateful for F. Gimbert, C. Stanciu and anonymous reviewer, for their helpful comments that contributed to improve this manuscript.

DATA AVAILABILITY

Data from the long-term regional seismic network JURAQUAKE and the JURASSIC KARST hydrogeological observatory of the SNO KARST network were used in the creation of this manuscript. A description of the data and their access are found in https://www.fdsn.org/networks/detail/5C_2018/ for JURAQUAKE and https://sokarst.org/en/9-observatories-in-france/jurassic-karst-en/ for JURASSIC KARST. Figures were made with Matplotlib version 3.5, available under the Matplotlib license at https://matplotlib.org/. The map was created through QGIS3 available under https://www.qgis.org/fr/site/forusers/download.html. The regression is done using the scikit learn library "RandomForestRegressor". The Python code for the computation of the features is created by C. Hibert and modified by A. Abi Nader for this article and can be found under https://doi.org/10.5281/zenodo.6592237.

REFERENCES

Andreo, B. *et al.*, 2006. Karst groundwater protection: first application of a pan-european approach to vulnerability, hazard and risk mapping in the sierra de líbar (southern spain), *Sci. Total Environ.,* **357** . doi:10.1016/j.scitotenv.2005.05.019.

Anthony, R.E., Aster, R.C., Ryan, S., Rathburn, S. & Baker, M.G., 2018. Measuring mountain river discharge using seismographs emplaced within the hyporheic zone, *J. Geophys. Res.: Earth Surf.,* **123** . doi:10.1002/2017JF004295.

Baillard, C., Crawford, W.C., Ballu, V., Hibert, C. & Mangeney, A., 2013. An automatic kurtosis-based P- and S-phase picker designed for local seismic networks, *Bull. Seism. Soc. Am.,* **104**(1), 394–409.

Breiman, L., 2001. Random forests, *Mach. Learn.,* **45,** 5–32.

Burtin, A., Bollinger, L., Vergne, J., Cattin, R. & Nábělek, J.L., 2008. Spectral analysis of seismic noise induced by rivers: a new tool to monitor spatiotemporal changes in stream hydrodynamics, *J. Geophys. Res.: Solid Earth,* **113** . doi:10.1029/2007JB005034.

Chen, Z. *et al.*, 2017. The world karst aquifer mapping project: concept, mapping procedure and map of europe, *Hydrogeol. J.,* **25**(3).

Chmiel, M., Walter, F., Wenner, M., Zhang, Z., McArdell, B.W. & Hibert, C., 2021. Machine learning improves debris flow warning, *Geophys. Res. Lett.,* **48**(3). doi:10.1029/2020GL090874.

Cholet, C., Steinmann, M., Charlier, J.-B. & Denimal, S., 2015. Comparative study of the physicochemical response of two karst systems during contrasting flood events in the french jura mountains, *Environ. Earth Sci.,* **1,** 1–9.

Cholet, C., Charlier, J.-B., Moussa, R., Steinmann, M. & Denimal, S., 2017. Assessing lateral flows and solute transport during floods in a conduit-flow-dominated karst system using the inverse problem for the advection-diffusion equation, *Hydrol. Earth Syst. Sci.,* **21**(7), 3635–3653.

Criminisi, A., Konukoglu, E. & Shotton, J., 2011. *Decision forests for classification, regression, density estimation, manifold learning and semi-supervised learning,* NOW Publishers, foundations and trends® in computer graphics and vision:**7**:2–3.

Díaz, J., Ruíz, M., Crescentini, L., Amoruso, A. & Gallart, J., 2014. Seismic monitoring of an Alpine mountain river,*J. Geophys. Res. Solid Earth,* **119**(4), 3276–3289. doi: 10.1002/2014JB010955.

Dietze, M., Lagarde, S., Halfi, E., Laronne, J.B. & Turowski, J.M., 2019. Joint sensing of bedload flux and water depth by seismic data inversion, *Water Resources Res.,* **55**(11), 9892–9904.

Drew, D., 1999. *Karst Hydrogeology and Human Activities: Impacts, Consequences and Implications: IAH International Contributions to Hydrogeology 20 (1st ed.).,* Routledge. doi:10.1201/9780203749692

Ford, D. C. & Williams, P., 2007. *Karst Hydrogeology and Geomorphology,* John Wiley & Sons.

Fores, B., Champollion, C., Mainsant, G., Albaric, J. & Fort, A., 2018. Monitoring saturation changes with ambient seismic noise and gravimetry in a karst environment, *Vadose Zone J.,* **17**(1), 170163.

Gimbert, F., Tsai, V.C. & Lamb, M.P., 2014. A physical model for seismic noise generation by turbulent flow in rivers, *J. Geophys. Res. (Earth Surf.),* **119**(10). doi:10.1002/2014JF003201.

Green, T.R., Taniguchi, M., Kooi, H., Gurdak, J.J., Allen, D.M., Hiscock, K.M., Treidel, H. & Aureli, A., 2011. Beneath the surface of global change: impacts of climate change on groundwater, *J. Hydrol.,* **405**(3), 532–560.

Hibert, C., Provost, F., Malet, J.-P., Maggi, A., Stumpf, A. & Ferrazzini, V., 2017. Automatic identification of rockfalls and volcano-tectonic earthquakes at the Piton de la Fournaise volcano using a Random Forest algorithm, *J. Volcanol. Geotherm. Res.,* **340,** 130–142.

Jourde, H. *et al.*, 2018. Sno karst: a French network of observatories for the multidisciplinary study of critical zone processes in karst watersheds and aquifers, *Vadose Zone Journal,* **17** . doi:10.2136/vzj2018.04.0094.

Kim, Y., Hardisty, R., Torres Parada, E. & Marfurt, K., 2018, Seismic-facies classification using random forest algorithm, *SEG Technical Program Expanded Abstracts,* pp2161–2165. doi:10.1190/segam2018-2998553.1.

Küperkoch, L., Meier, T., Lee, J., Friederich, W. & Group, E.W., 2010. Automated determination of P-phase arrival times at regional and local distances using higher order statistics, *J. geophys. Int.,* **181**(2), 1159–1170.

Larose, E. *et al.*, 2015. Environmental seismology: what can we learn on earth surface processes with ambient noise?, *Journal of Applied Geophysics,* **116,** 62–74. doi:10.1016/j.jappgeo.2015.02.001.

Liang, Z., Wei, J., Zhao, J., Liu, H., Li, B., Shen, J. & Zheng, C., 2008. The statistical meaning of kurtosis and its new application to identification of persons based on seismic signals, *Sensors (Basel, Switzerland),* **8**(8), 5106–5119.

Ma, M., Wang, S., Yuan, S., Wang, J. & Wang, T., 2015. The comparison of skewness and kurtosis criteria for wavelet phase estimation, *SEG Technical Program Expanded Abstracts 2015*. doi:10.1190/segam2015-5826646.1.

McCuen, R.H., Knight, Z. & Cutter, A.G., 2006. Evaluation of the Nash-Sutcliffe efficiency index, *J. Hydrol. Eng.,* **11**(6), 597–602

McDonnell, J.J., 2017. Beyond the water balance, *Nature Geosci,* **10**(6), 396–396. doi:10.1038/ngeo2964

Provost, F., Hibert, C. & Malet, J.-P., 2017. Automatic classification of endogenous landslide seismicity using the random forest supervised classifier, *Geophys. Res. Lett.,* **44**(1), 113–120.

Ross, Z.E. & Ben-Zion, Y., 2014. Automatic picking of direct P, S seismic phases and fault zone head waves, *J. geophys. Int.,* **199**(1), 368–381.

Rouet-Leduc, B., Hulbert, C., Lubbers, N., Barros, K., Humphreys, C.J. & Johnson, P.A., 2017. Machine learning predicts laboratory earthquakes, *Geophys. Res. Lett.,* **44**(18), 9276–9282.

Schmandt, B., Aster, R.C., Scherler, D., Tsai, V.C. & Karlstrom, K., 2013. Multiple fluvial processes detected by riverside seismic and infrasound monitoring of a controlled flood in the grand canyon, *Geophys. Res. Lett.,* **40**(18), 4858–4863.

Tsai, V., Minchew, B., Lamb, M. & Ampuero, J.P., 2012. A physical model for seismic noise generation from sediment transport in rivers, *Geophys. Res. Lett.,* **39,** L02404. doi:10.1029/2011GL050255.

Vidal, C., Zaccarelli, L., Pintori, F., Bragato, P. & Serpelloni, E., 2021. Hydrological effects on seismic-noise monitoring in karstic media, *Geophys. Res. Lett.,* **48,** e2021GL093191. doi:10.1029/2021GL093191.

Voisin, C., Guzmán, M.A.R., Réfloch, A., Taruselli, M. & Garambois, S., 2017. Groundwater monitoring with passive seismic interferometry, *J. Water Resour. Prot.,* **9**(12), 1414–1427.

Wenner, M., Hibert, C., van Herwijnen, A., Meier, L. & Walter, F., 2021. Near-real-time automated classification of seismic signals of slope failures with continuous random forests, *Nat. Hazards Earth Syst. Sci.,* **21**(1) . doi:10.5194/nhess-21-339-2021.

Zou, C., Zhao, L., Xu, M., Chen, Y. & Geng, J., 2021. Porosity prediction with uncertainty quantification from multiple seismic attributes using random forest, *J. Geophys. Res.: Solid Earth,* **126**(7) e2021JB021826. doi:10.1029/2021JB021826.

APPENDIX A: SPECTROGRAMS OF THE HORIZONTAL COMPONENTS OF THE SEISMIC SIGNALS

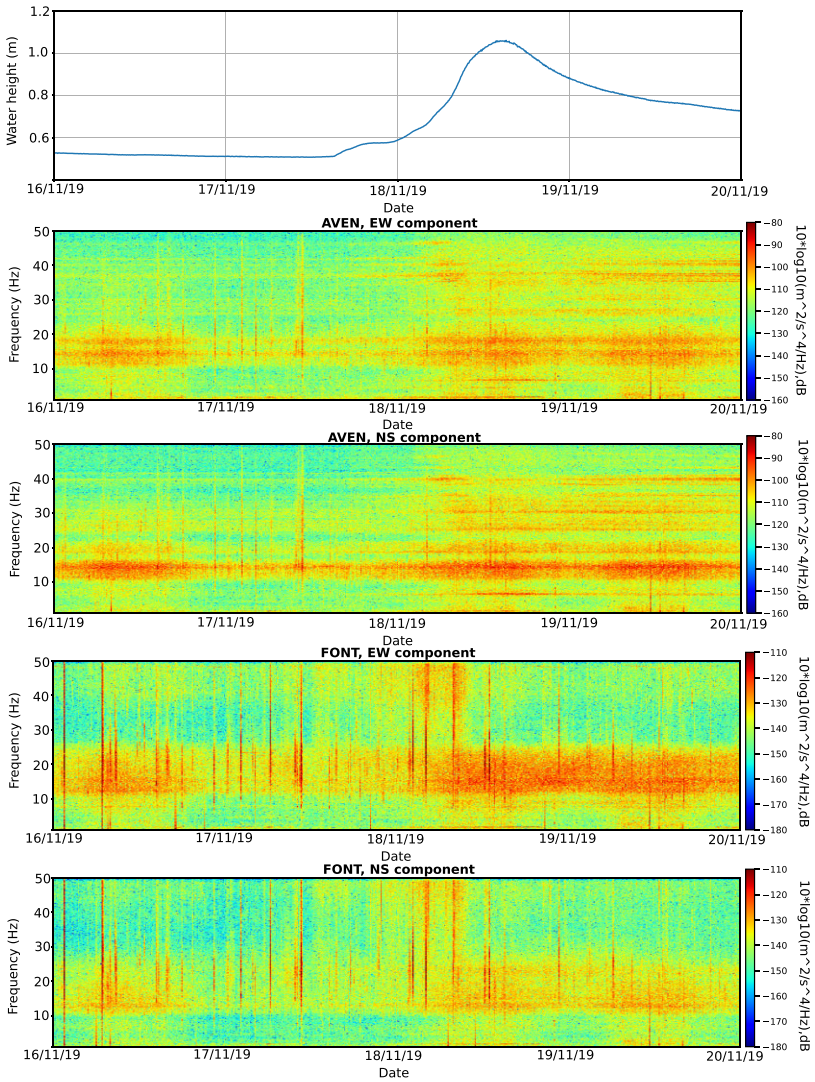

Figure A1. Water height for the flood recorded between 16/11/19 and 20/11/19 along with the spectrograms of the horizontal components during the selected flood for signals recorded at AVEN and FONT and filtered between 1 and 50 Hz. EW corresponds to the East-West component, NS to the North-South component.

APPENDIX B: TEMPORAL EVOLUTION OF FEATURES

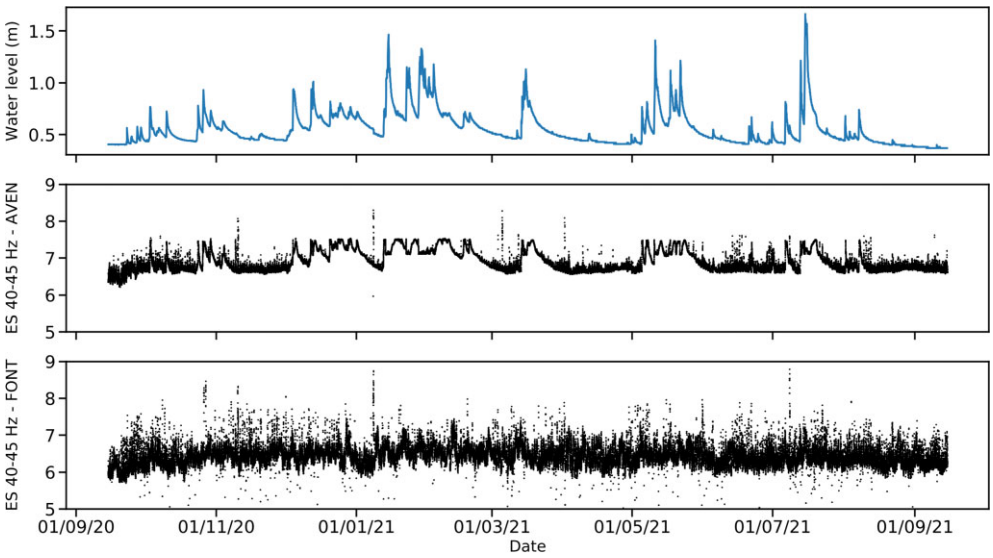

Figure B1. Water height recorded for the hydrological cycle used for the training and the signal's energy feature between 40 and 45 Hz for the signals recorded at AVEN and FONT during the same cycle.

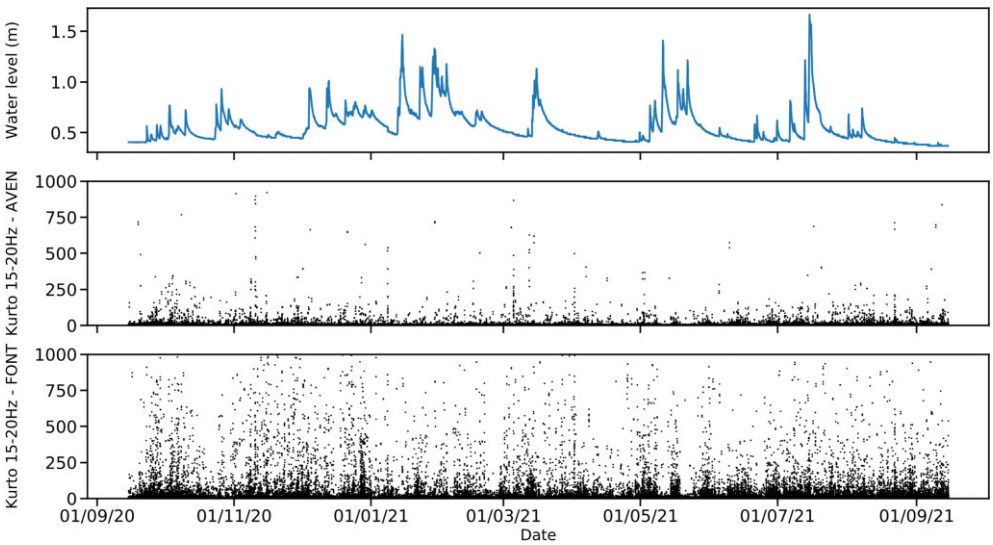

Figure B2. Water height recorded for the hydrological cycle used for the training and the signal's kurtosis feature between 15 and 20 Hz for the signals recorded at AVEN and FONT during the same cycle.