What's Behind This Water Table Depth Forecasting? RISE Application for Spatial, Temporal, and Spatio-Temporal Explanations

Matteo Salis[1,3,*], Gabriele Sartor[1], Marco Pellegrino[1], Stefano Ferraris[2], Abdourrahmane M. Atto[3] and Rosa Meo[1]

[1]*Computer Science Department - University of Turin, Corso Svizzera 185, Torino Italy*

[2]*Interuniversity Department of Regional and Urban Studies and Planning - Politecnico di Torino and University of Turin, Viale Pier Andrea Mattioli 39, Torino Italy*

[3]*LISTIC Laboratory - Université Savoie Mont Blanc, 5 chemin de bellevue, 74 940 Annecy-le-vieux France*

Abstract

Over the past few years, the effects of climate change have significantly increased, directing our attention to our environmental resources. One of the most critical resources is fresh water, whose availability has been endangered by even more frequent extreme events of precipitations and droughts. Deep Learning (DL) has revealed a useful tool for accurate hydrological forecasting, but its main drawback is its black-box nature. To face this issue eXplainable Artificial Intelligence (XAI) has come out. In this work, Randomized Input Sampling for Explanation (RISE) is applied to a regression spatio-temporal model trained to predict the Water Table Depth (WTD) collected by a sensor in Vottignasco, in the northwest of Italy. An investigation of the model behaviour over spatial, temporal and spatio-temporal dimensions has been conducted formalizing S-RISE, T-RISE, and ST-RISE respectively. Results suggest the usefulness of a spatio-temporal approach (ST-RISE) and give interesting intuitions on the events that occurred over the area.

Keywords

Explainable AI, Model agnostic algorithms, RISE, Spatio-temporal explanations,

1. Introduction

Water is essential to all species. Groundwater resources represent one of the most relevant sources of freshwater for human beings. Changes in rainfall patterns and temperature due to climate change have made groundwater even more pivotal in providing fresh and clean water to communities [2, 3, 4]. Groundwater resources could be quantified by measuring the water table depth (WTD), i.e. the distance between the ground surface and the higher surface of the groundwater body. Numerical models have been successfully proposed to model the WTD, however, they usually depend heavily on the geophysical properties of the area to be modelled. This requires extensive measurements and computationally intensive simulations. Deep Learning (DL) models have been adopted to overcome this issue and develop models which depend only on open and easy-to-measure weather data (e.g. rainfall and temperature). More specifically, spatio-temporal DL models based on computer vision architectures have been proposed to handle dynamical relations between weather and the hydrogeological output [5, 6, 7]. For this model type the input is often a weather video, in which each frame τ contains a weather map. Each weather map comprises multiple channels C, in other words, the observed weather variables (e.g. rainfall, temperature etc.). The weather video is thus fed into a DL model to forecast a specific hydrological variable (e.g. the WTD, see Figure 1).

AI4CC-IPS-RCRA-SPIRIT 2024: International Workshop on Artificial Intelligence for Climate Change, Italian Workshop on Planning and Scheduling, RCRA Workshop on Experimental evaluation of algorithms for solving problems with combinatorial explosion, and SPIRIT Workshop on Strategies, Prediction, Interaction, and Reasoning in Italy. November 25-28th, 2024, Bolzano, Italy [1].

*Corresponding author.

matteo.salis@unito.it (M. Salis)

https://sites.google.com/view/matteo-salis/home (M. Salis)

0009-0009-2810-9992 (M. Salis); 0000-0002-6530-318X (G. Sartor); 0000-0003-1753-4917 (A. M. Atto)

DL models achieved overwhelming performances [8, 9, 10, 7]; however, their black box nature is a matter of fierce debate in the scientific community and prevents their adoption as a reliable tool to support decision and policymaking. An entire research field named explainable Artificial Intelligence (XAI) has come out with the aim of explaining DL models behaviour. There exist several approaches which depend on the generalization with respect to DL architecture (*model specific* or *model agnostic*) and to data (*global explanation* or *local explanation*)[11, 12, 13]. More specifically, an algorithm is said to be *model agnostic* if it could be applied regardless of the specific type of DL architecture. Concerning data generalization, an explanation is defined *local* if it is specific to a dataset instance, reversely it is named *general* when it is related to the overall behaviour of the model over the whole dataset. Many strategies have been proposed to produce explanations, some of them, like gradient-based and propagation-based techniques, require access to some or all hidden layers. Reversely, perturbation-based techniques, among which the local method RISE (Randomized Input Sampling for Explanation) [14], need only a trained model by which to perform predictions. RISE was proposed for image classification explanations, and it consists of masking J times randomly the input features of the instance to be explained. Then, a saliency map is created highlighting pixels that contribute the most to the class probability score maximization. This approach is rather simple and suitable when details about the black box model are not reported and one cannot access hidden activations.

The main objective of this study is to produce local explanations for a specific CNN-LSTM black box model proposed in [7] which forecasts the weekly WTD given a weekly weather video of the previous two years of total precipitation, minimum temperature, and maximum temperature. Given the multidimensionality (time and space) of the input features applying a XAI algorithm is not a trivial task, especially because many of them have been originally developed for standard classification tasks on well-known public dataset [15, 14, 16, 17]. More specifically, in the presence of spatio-temporal data (e.g. weather video), it is relevant to produce not only a spatial saliency map, but also a temporal, and even more importantly a spatio-temporal saliency object for each input feature (i.e. weather variable or video's channel). To this aim, we have adopted the perturbation-based and *model-agnostic* RISE algorithm, and we have proposed and formalized its application to produce **a)** a spatial explanation in the form of a saliency 2D map (very similar to the RISE original version) which shows the most relevant pixels; **b)** a temporal explanation in the form of a saliency 1D vector which highlights the most relevant frames in the input video; and **c)** a spatio-temporal explanation in the form of a saliency 3D video which highlights the most relevant pixels for each frame. Given the preliminary nature of the present study, we have just focused on total precipitations, which is also the most interesting weather variable for domain experts. The choice of RISE instead of other XAI methods is because of its simplicity and because the extension to multidimensional data (e.g. video) has been already reported as a future work by authors in the original paper [14]. Furthermore, RISE has been already applied in hydrological DL studies [18, 5] but without making any extension and formalization to temporal and spatio-temporal explanations.

2. Related works

Some XAI algorithms have already been developed for video-related tasks. Authors in [19] proposed saliency Tubes for highlighting the most relevant areas in each frame of a video for a classification task. However, this *model-specific* method has been developed for 3D CNN and requires access to the last convolutional layer activations. [20] proposed a new, but still perturbation-based, method named STEP (Spatio-Temporal Extremal Perturbation) which finds the salient elements using 3D masks generated via a 2-step optimization procedure. In more detail, STEP finds the smallest subset of input elements for retaining the highest prediction accuracy on a specified target label. The downside of this more sophisticated approach is that it introduces new hyperparameters, another loss function to be chosen, and a new heavy optimization problem to be solved. Authors in [21] criticized the adoption of a 3D mask in case of complex action recognition tasks. Thus, they proposed to generate masks based on optical flows which estimate apparent object motions inside a video. However, these optical flow need

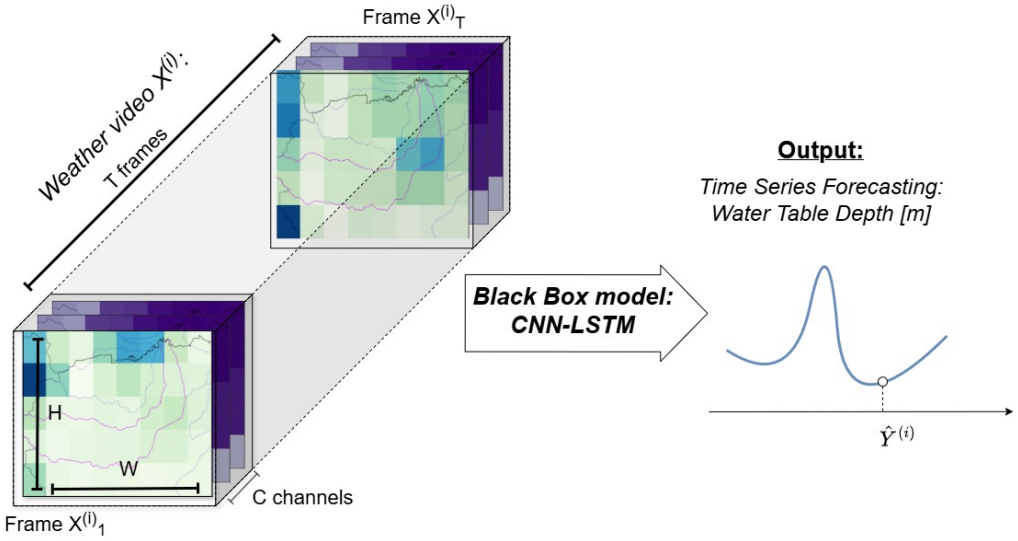

Figure 1: Model's input-output: a weather video (composed by the three channels precipitations, minimum and maximum temperatures) are given as input to the CNN-LSTM model of [7] returning the predicted Water Table Depth in meters for a specific station.

to be in turn estimated with other algorithms.

It is relevant to stress that our explanation task should be simpler than video classification tasks in which high-resolution video contains complex actions to be recognized. Indeed, in the case study of [7] the images are about 12.5km in resolution and the task is to forecast a scalar, not recognize and classify a complex motion. For these reasons, even if STEP and Adaptive Occlusion could be very suited for some more complex case studies, we decided to adopt the simpler and *model-agnostic* RISE algorithm.

3. Methods

As already stated, RISE has been proposed to retrieve pixel relevance in a classification task, detecting the pixels in the input images that maximize the class probability scores. The result of this method is a spatial explanation, i.e. a saliency map over the input's pixels, in which the higher the value, the higher the pixel's influence is on the model's prediction. In the original paper [14] input perturbations were performed by applying black blurred patches over the unperturbed input. Our case study has required some modifications. The first major modification is because we are dealing with a regression task. Thus, inspired by [18, 12, 5], we have defined a saliency object for the i-th instance after J perturbations as in Equation 1:

$$S^{(i)} = \frac{1}{\overline{M}^{(i)} \cdot J} \sum_{j=1}^{J} \Delta_j^{(i)} \cdot M_j^{(i)}, \tag{1}$$

where $\overline{M}^{(i)}$ is the masks' mean. In other words, it is a weighted sum of random masks, where the weights Δ_j^i (Equation 2) are the absolute difference between the unperturbed prediction $\hat{Y}^i$ and the perturbed prediction $\hat{Y}_{p_j}^i$ obtained respectively by feeding the model with the original input $X^{(i)}$ and the input perturbed with the j-th perturbation mask.

$$\Delta_j^{(i)} = |\hat{Y}^{(i)} - \hat{Y}_{p_j}^{(i)}|, \tag{2}$$

The second major modification is related to the physical meaning of data contained in the input weather video. Indeed, whilst in the case of classical RGB image blurring with zeros means inserting black smooth patches (i.e. no information), in the case of physical raster data substitute values with zeros has a strong physical meaning. To this aim, inspired by [22, 23, 18, 5], we decide to perturb input

images with additive Gaussian noise. In this way, when a positive noise is generated the precipitations are increased, reversely if the negative noise is generated the precipitations are decreased.

Given the multidimensional nature of the input, perturbation masks with different dimensions could be created. This will bring saliency objects with different dimensions and meanings. Following this idea and to accomplish our aim to define a spatial, a temporal, and a spatio-temporal explanation, we have formalized three different RISE applications varying the dimensionality of the Gaussian noise that defines the perturbations:

1. **S-RISE**: In this case, noise is in the form of 2D perturbation mask. For every iteration j a 2D mask M_j is applied to each frame of the input video. The result is a 2D saliency map, similar to the original RISE, which highlights the areas that are more influential on the model's i-th prediction.
2. **T-RISE**: In this case, noise is in the form of 1D perturbation mask. In every iteration j a perturbation vector M_j is a univariate Gaussian noise, and it is applied to each pixel time series (i.e. the series of values of a pixel over the whole video). The resulting saliency object is a vector which determines the most relevant frame in the input video for the i-th prediction.
3. **ST-RISE**: In this last formalization, noise is in the form of a multivariate 3D Gaussian perturbation mask, i.e. every mask M_j is a video of the same spatial and temporal extent as the input one. Then, the noise is applied element-wise to the unperturbed input. Thus, the result is a saliency video which highlights the most relevant areas and frames for the i-th prediction.

In the following subsections, each approach is formalized.

3.1. S-RISE

In our work, to predict a groundwater level $Y^{(i)} \in \mathbb{R}$ we consider a set of N videos $X = \{X^{(i)} | i \in [1; N]\}$. A video is a tensor $X^{(i)}$ with element $X^{(i)}_{h,w,c,\tau} \in \mathbb{R}$ representing the pixel at height $h \in [1, ..., H]$, width $w \in [1, ..., W]$, channel $c \in [1, ..., C]$ and number of frame $\tau \in [1, ..., T]$. Then, the model used in this work is defined as $f : X \rightarrow Y$, which returns the t-th weekly WTD defined as $\hat{Y}^{(i)}$, taking as input a video $X^{(i)}$, representing the weather video of the T previous weeks before the target dates.

As previously said, S-RISE consists in calculating $\Delta^{(i)}_j$ for each input sample $X^{(i)}$, for J iterations. In this case, $\hat{Y}^{(i)}_{p_j}$ is the prediction $f(X^{(i)} + M_j)$ produced by processing $X^{(i)}$ perturbed by the two-dimensional Gaussian noise M_j. Specifically, in the spatial case, the two-dimensional noise of the iteration j for the instance i is defined as:

$$M_j = \beta \cdot e^{-\frac{1}{2}\left[\frac{(h-h_p)^2}{\sigma^2} + \frac{(w-w_p)^2}{\sigma^2}\right]}, \tag{3}$$

where $\beta \in \{-1, 1\}$ is a random parameter to make positive or negative the perturbations, (h_p, w_p) the coordinates of the Gaussian noise centre which are randomly sampled from a uniform distribution, and σ^2 defines the extension of the Gaussian noise in the space. Formally, in the spatial case, the perturbation of the input over a specific channel c can be defined as

$$X^{(i)} + M_j = X^{(i)}_{c,\tau} + M_j, \tag{4}$$

$\forall \tau \in [1, ..., T]$ and a specific channel $c \in [1, ..., C]$, conceptually meaning adding the 2D noise map M_j to each frame τ on the channel c.

Finally, the 2D saliency map for the i-th instance is obtained from Equation 1 after J iterations (see pipeline on the top of Figure 2).

3.2. T-RISE

In the T-RISE an input video $X^{(i)}$ is perturbed over the temporal dimension, i.e. along the frames. For this purpose, a one-dimensional perturbation mask M_j is generated at each iteration j as a univariate Gaussian noise defined as follows:

$$M_j = \beta \cdot e^{-\frac{1}{2}\left[\frac{\tau - t_p}{\sigma^2}\right]}, \tag{5}$$

with the centre of the disturbance in the frame t_p that is randomly sampled from a uniform distribution. In the temporal case, the perturbation of the input over a specific channel c can be defined as:

$$X^{(i)} + M_j = X^{(i)}_{h,w,c} + M_j, \tag{6}$$

$\forall h \in [1, ..., H]$, $\forall w \in [1, ..., W]$ and a specific channel $c \in [1, ..., C]$, producing a new video applying the noise vector over all the pixel time series $X^{(i)}_{h,w,c}$ in the same way. In this case, $S^{(i)}$ captures the influence of each frame (i.e. time step) on the prediction and, consequently, it is represented as a saliency vector of dimension T (see pipeline at the bottom of Figure 3).

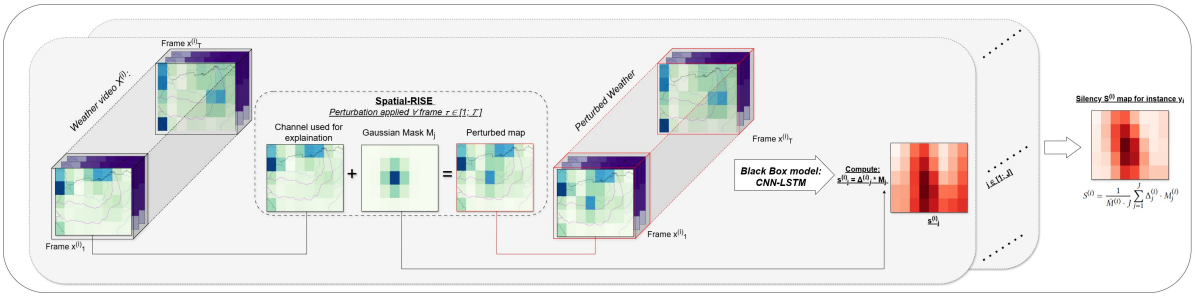

Figure 2: Pipeline for S-RISE perturbing the precipitations of the input weather video and producing a 2D saliency map.

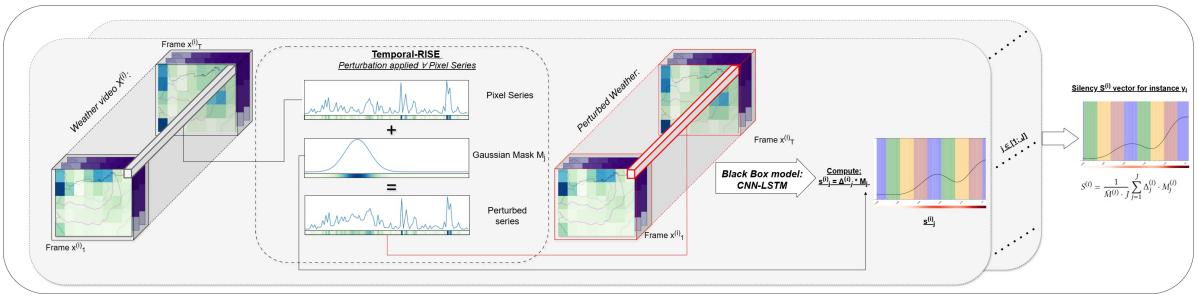

Figure 3: Pipeline for T-RISE perturbing input's precipitations over the temporal dimension (all the pixels of a specific frame (i.e. time-step) are perturbed by the same amount of noise) and generating 1D saliency vector.

3.3. ST-RISE

In the last case, to assess the joint influence of spatial and temporal dimensions, RISE is applied using a three-dimensional Gaussian noise defined as:

$$M_j = \beta \cdot \exp\{-\frac{1}{2}(\mathbf{x} - \mu)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x} - \mu)\}, \tag{7}$$

where μ are the coordinates in space and time of the pixel which is the centre of the Gaussian noise at iteration j and are randomly sampled from uniform distributions. Finally, the perturbation of the input in the spatio-temporal approach can be seen as

$$X^{(i)} + M_j = X^{(i)}_{c'} + M_j, \tag{8}$$

representing the addition of the noise video M_j to the complete weather video for a specific channel $c \in [1, ..., C]$. Consequently, the saliency video $S^{(i)}$ identifies the most influential pixels' value considering jointly time and space (see pipeline in Figure 4).

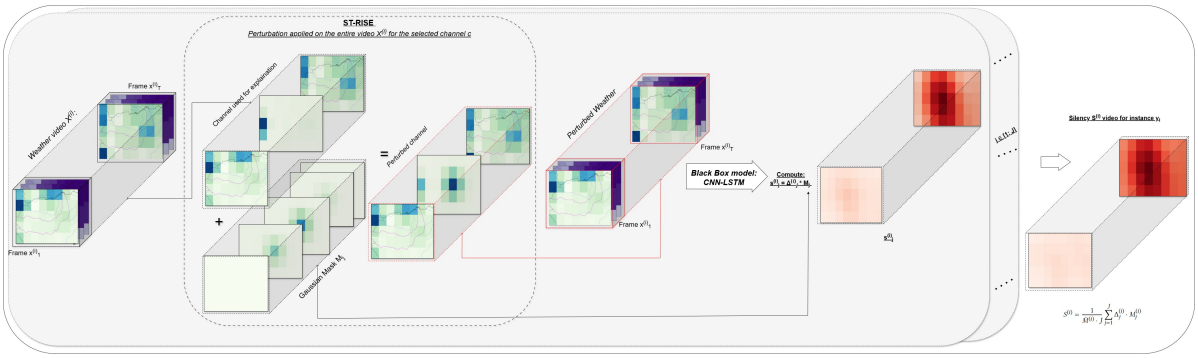

Figure 4: Pipeline for ST-RISE, differing from the previous cases for using multivariate 3D Gaussian noise, explicating the most influential pixels jointly over time and space.

4. Results and discussions

4.1. Data

In [7] authors defined their Region of Interest (ROI) as the Grana-Maira catchment in Piemonte, an Italian administrative region (Figure 5). Weather videos were set up using total precipitations (both snow and rain), minimum temperature and maximum temperature, i.e. $C = 3$ channels. In the original paper, WTD time series data were retrieved for different sensors located in the same catchment, and for each sensor, a local model was trained. However, because of our interest in method formalization, at this stage of the research, we have decided to focus on just one sensor in the municipality of Vottignasco. Authors made available their data already pre-processed, standardized by computing z-scores and split into training, validation, and test. We decided to focus only on the standardized test instances, looking at the model behaviour in a scenario mostly independent from training data.

4.2. Implementation details

We have experimented with the three proposed approaches S-RISE, T-RISE, and ST-RISE for each one of the 104 instances present in the test set of the Vottignasco WTD series. For each approach, β has been sampled from a discrete uniform distribution with values 1 and -1. For S-RISE σ has been set to 0.75 that, given the resolution of weather data, corresponds approximately to 10km. Instead for T-RISE σ has been set to 8, which corresponds to two months, thus roughly 95% of the noise is concentrated in four months. For both S-RISE and T-RISE the total number of iterations J to produce a saliency object has been set to 1000. This was found, inspired by [5], by looking for the minimum number of iterations that makes the saliency object vary less than the 2% from further perturbations. For ST-RISE the covariance matrix Σ has been constructed by setting covariances to 0 and using the individual standard deviations of the previous cases (0.75 for spatial dimensions and 8 for the temporal one). The total number of iterations J in the case of ST-RISE has been increased to 5000. All experimentation has been conducted on Google Colab using the free available resources[1].

4.3. Evaluation

The application of RISE in its three variations is assessed through insertion/deletion [14] metrics and visual inspection of the saliency representations validated by domain experts. In the first case, using some metrics we have a numerical and more objective evaluation of the different approaches while, in the latter, saliency representations can give a more intuitive explanation of the model's behaviour regarding the dimension on which RISE focuses on. More in detail, the insertion consists of iteratively calculating for each instance i the error between the original model's prediction $\hat{Y}^{(i)}$ and the one "enabling" only a percentage k of most important pixels of the input, according to the obtained saliency

[1]The implementation of the experiments is available at this link.

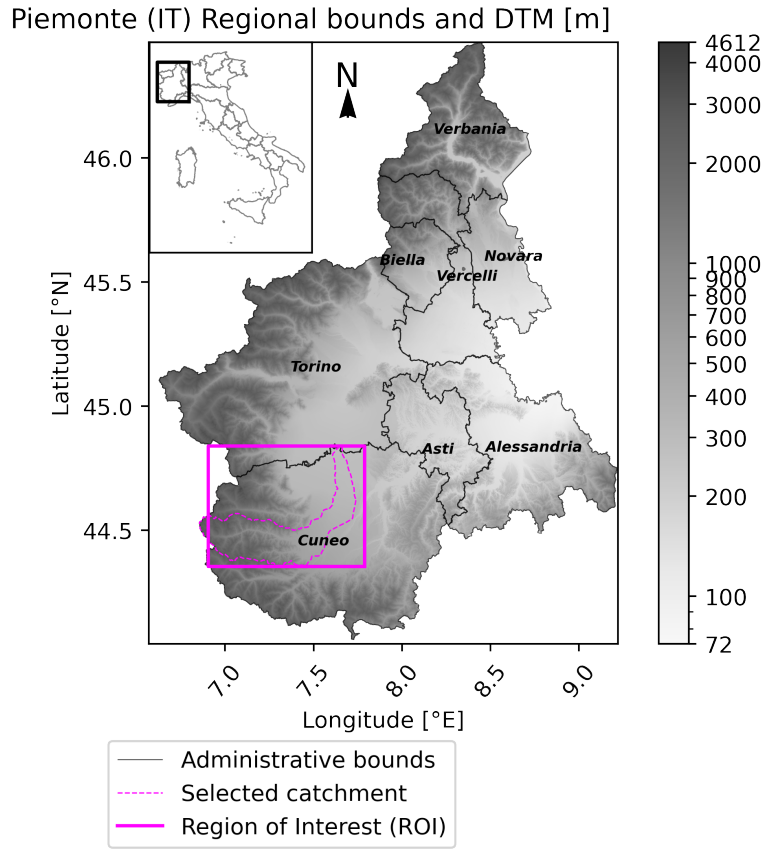

Figure 5: ROI: Grana-Maira catchment in Piemonte (IT). Figure 1 in [7].

representations, leaving all other pixels to 0. Reversely, the deletion starts from the original input and iteratively deletes information zeroing out the k percent of the most relevant input pixels. Therefore, Insertion and deletion metrics present complementary assessments.

4.3.1. Insertion & Deletion

Given our regression task, both insertion and deletion are computed by looking at the mean squared error between the original unperturbed prediction and the one obtained by feeding the inserted/deleted perturbed input to the model.

It is relevant to emphasize that S-RISE, T-RISE, and ST-RISE produce saliency objects of different dimensions. The saliency video of ST-RISE is the most precise, and it allows sorting each element (tuple of pixel and corresponding frames) of the input video by its relevance to the prediction. Differently, S-RISE and T-RISE could produce rankings only in space and time respectively. This means that for each deletion or insertion step for S-RISE T elements (i.e., a pixel time series) are zero-out or enabled, while for T-RISE $H * W$ elements are set to 0 or enabled. The saliency video of ST-RISE is the only one that allows the computation of insertion and deletion metrics for each single element of the input video.

As already discussed in Section 3, another specificity of our case study is that setting pixels to 0 has a physical meaning. Given that our dataset is standardized and feature means' are then $\mu_c = 0$ $\forall c \in [1, ..., C]$ where $C = 3$, in the case of deletion, we are eliminating relevant information by replacing the mean; while insertion starts with a video made of a constant value (i.e. 0, the feature's mean) and at each iteration information is restored.

On the left of Figure 6 it is depicted the mean of the insertion curves computed on all the 104 instances of the test set. On the x-axis is reported the fraction of input elements (i.e., tuple of pixel and frame) enabled. The idea behind the insertion metric is to evaluate how much the most influential pixels

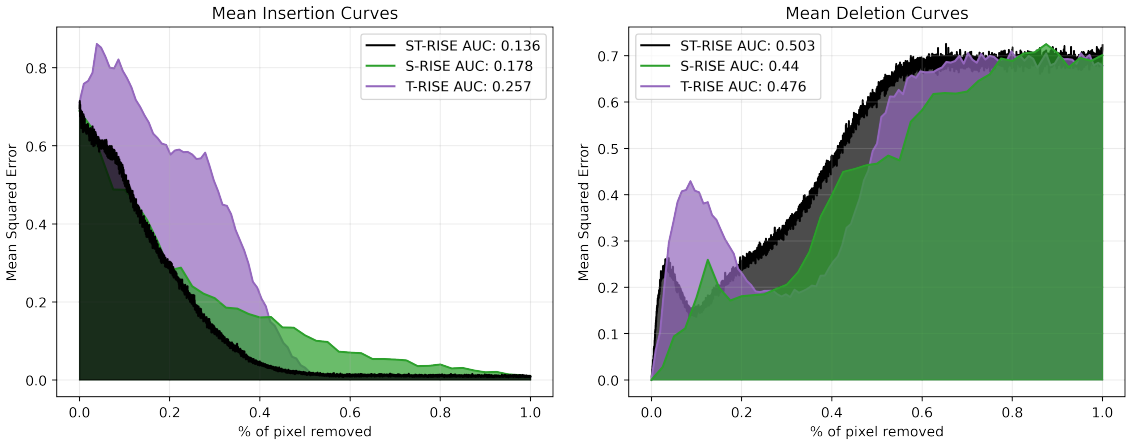

Figure 6: Analysis of the most influential pixels on the model's prediction using insertion (left) and deletion (right) metrics. The metrics are calculated as mean over all the instances, respectively, inserting/removing the same portion of most influential pixels to each sample for S-RISE (green), T-RISE (purple) and ST-RISE (black).

influence the prediction without nearby information. The sharpest is the drop in the metric, the higher is the relevance of that portion of input [14]. Consequently, a lower AUC (Area Under the Curve) means a better explanation which isolates in a more precise way the most relevant pixels. ST-RISE has revealed the best achieving an AUC of 0.136. This means that ST-RISE has detected more precisely the most relevant portions of the input on average on all the test instances. Of course, this result is thanks to the higher definition of the saliency video with respect to the saliency map and saliency vector. Furthermore, ST-RISE is in principle able to capture jointly spatial and temporal (i.e. spatio-temporal) relations, while S-RISE and T-RISE are only able to look for explanations in spatial and temporal dimensions separately.

Figure 6 depicts on the right the mean deletion curve computed averaging the single deletion curve over the 104 test instances. It represents the increase in the error, deleting a percentage k of the most influential pixels from the original output. In contrast to the insertion, the deletion measures the relevance of a limited number of pixels keeping the others unchanged (i.e. maintaining contextual information). The higher the increase in the curve, the better the explanation is. In this case, ST-RISE confirms its higher effectiveness and S-RISE performs slightly worse than T-RISE with an AUC of, respectively, 0.503, 0.44 and 0.476. All the curves after a first peak display a consistent drop in the error and then all of them recover. This could be due to hallucination effects created by replacing true input values with a constant (the mean) value [14, 24, 25].

4.3.2. Graphical Explanation

In order to deduce a more conceptual explanation, S-RISE, T-RISE and ST-RISE are analyzed through visual inspection of their saliency representations. Indeed, Figure 7 shows the explanations generated by RISE on two instances of the test set. The first instance presents the saliencies for the prediction at the beginning of the spring 2022 (27-03-2022), while the latter one of the winter 2023[2] (25-12-2022).

From the spatial perspective, terrain's structure is pivotal in groundwater phenomena because water under the terrain flows from mountain to valley and then plain. This means that the WTD level is extremely related to the mountains and valleys' weather nearby the sensor location. Thus, given that in our ROI the highest mountains are in the northwest, we expect the model to be very sensible to precipitation in that area. The model seems to fulfil our expectations. Indeed, the most relevant pixels in S-RISE for both instances are focused on the west and northwest with respect to the sensor's location. Even if most relevant pixels are outside the catchment area, this explanation is plausible. Indeed, hydrological catchments are mainly defined focusing on river courses and thus groundwater bodies could be related to multiple catchments. Focusing on the Vottignasco WTD sensor, even if it is included

[2]We consider as winter 2023 the period of time between 21-12-2022 and 20-03-2023.

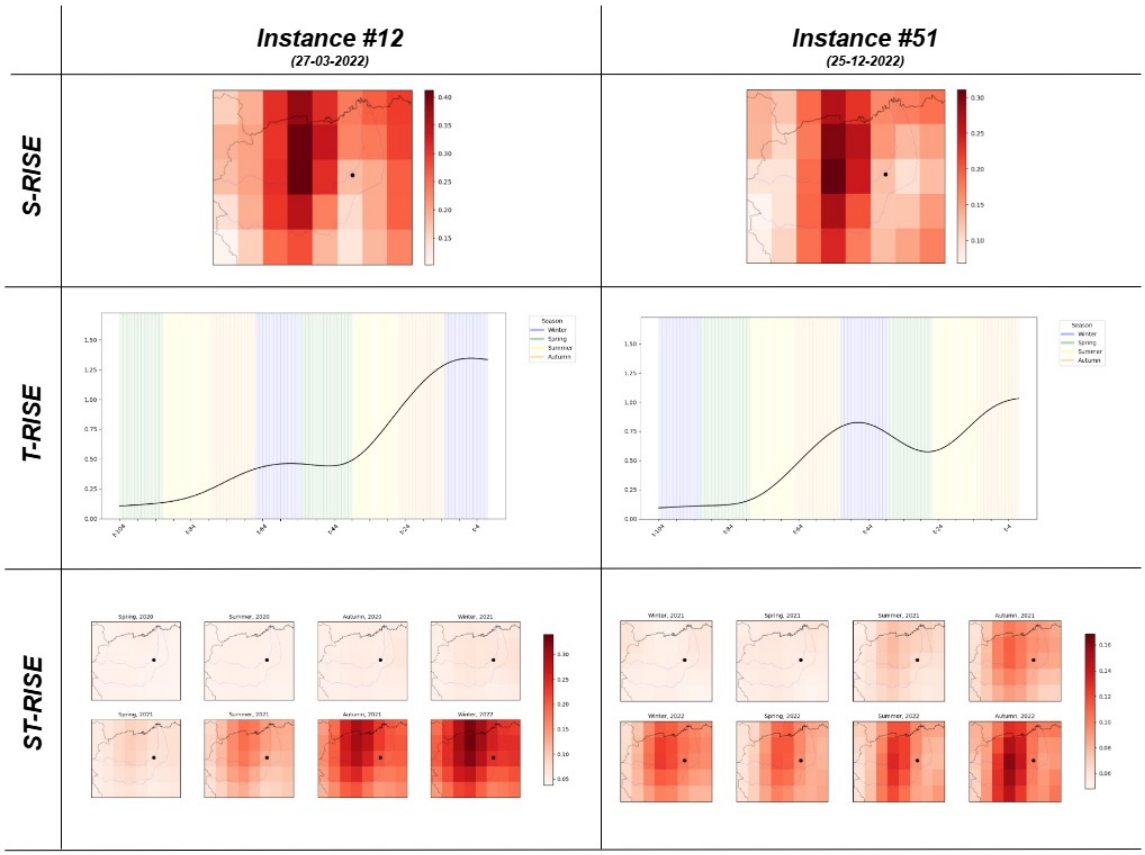

Figure 7: Comparison between of the instances 12 and 51 predicting, respectively, the WTD on 27-03-2022 and 25-12-2022. In the top row, we see the saliency map produced by S-RISE, placed on the administrative bounds map and highlighting the location of the Vottignasco's sensor (black dot). In the middle row, the saliency vector of T-RISE representing the influence of each week on the prediction at time t (from week $t - 104$ on the left, to week $t - 1$ on the right). Finally, the saliency video of ST-RISE is presented as the mean of the frames with respect to their season. The GIFs of the saliency video can be seen in the experiments repository.

in the Grana-Maira catchment, it is right in front of the Varaita valley[3], and from the domain knowledge it is sensible to have influences from even northern valleys. This is exactly what the saliency map from S-RISE highlights.

Concerning T-RISE, the saliency vectors produced for the two instances present substantial differences. The saliency vector of instance 12, inferring the WTD in spring, suggests that the prediction is mostly influenced by precipitation in winter 2022 and autumn 2021. A smaller impact on the model's output is also provided by data in winter 2021. Given that snow is most responsible for long-term dependencies and rain exerts mainly mid-short-term effects on groundwater bodies, it is sensible to interpret the saliency of winter 2021 as mainly driven by snow.

The saliency vector of instance 51 shows that the most influential precipitation period for the prediction in winter 2023 is the previous season, i.e. autumn 2022, but the former winter (2021) is almost equivalently relevant, highlighting a very long-term dependency. This could be because of the precipitation scarcity in 2022 and because at the beginning of winter 2023 the snow had been very rare and thus the previous year's snow-pack gained relevance.

The major benefit of using ST-RISE is that it is possible to look for spatio-temporal relevance patterns for each element of the input (i.e. pixel and frame tuples). For the instance 12, it appears that the most influential pixels of the input lie almost in the same locations over time. Consequently, no spatio-temporal events emerge from this representation which are not deducible from S-RISE and T-RISE disjointly. Instead, for the instance 51, a slight but visible shift of the saliency over the spatial and

[3]The Varaita catchment is just above in the north with respect to the Grana-Maira catchment.

temporal dimensions is present. More in detail, in winter 2022 the saliency is concentrated on the central northern area, while in autumn 2022 it has moved to the south, focusing on the central and the lower part of the ROI. This is in line with the possible interpretation of the temporal explanation given by the T-RISE saliency vector. Indeed, given the low precipitation in 2022 and low snow in early winter 2023, the previous year's snow-pack (winter 2022) could have been very relevant. It could be worth considering that in the northwest there is the Monviso, which is the highest mountain in the Cottian Alps, and thus, that area is a considerable snow reserve. Southern areas could have more relevance as time approaches the prediction date because of the rainfall collected by the Grana-Maira and Varaita catchment. Indeed, as already stated, the Vottignasco sensor is just in front of the Varaita Valley and it is very near the Maira River (which originates in the Grana-Maira catchment). This spatio-temporal saliency pattern could not have been detected by S-RISE and T-RISE separately, it is exactly for this reason that the usefulness of ST-RISE emerges.

5. Conclusions

In this paper, we have proposed a formalization of RISE application for spatial (S-RISE), temporal (T-RISE) and spatio-temporal (ST-RISE) explanations in a regression task. These implementations of RISE have been applied to the CNN-LSTM model described in [7], predicting the WTD in Vottignasco (Piemonte, IT) taking in input a stream of weather image data (i.e. a video). The saliency representations produced in this setting resulted explicative of the events occurring over time and space. In particular, ST-RISE has proved to be significantly useful in detecting jointly areas and times most relevant for a particular channel (i.e. variable) for predicting a specific instance. In other words, its local explanations are more fine-grained than the S-RISE and T-RISE, allowing recognition of saliency patterns that with S-RISE and T-RISE solely would not have been possible. Nonetheless, ST-RISE is compensated for its precision by incurring a higher computational cost, needing 5000 iterations instead of 1000 of the other approaches.

This work can be extended in multiple directions. First, in this study, we have just focused on one single variable. It could be worthwhile to investigate also the other available variables (i.e. channels) either individually or jointly. Moreover, a comparison with other XAI methods is needed to highlight the merits and limits of the presented approach, especially for spatio-temporal regression tasks.

RISE has proved to be useful in detecting the most relevant areas and times for WTD forecasting. In the end, this could be extremely helpful for environmental monitoring and policy implementation, fostering regional resilience and sustainability.

References

[1] D. Aineto, R. De Benedictis, M. Maratea, M. Mittelmann, G. Monaco, E. Scala, L. Serafini, I. Serina, F. Spegni, E. Tosello, A. Umbrico, M. Vallati (Eds.), Proceedings of the International Workshop on Artificial Intelligence for Climate Change, the Italian workshop on Planning and Scheduling, the RCRA Workshop on Experimental evaluation of algorithms for solving problems with combinatorial explosion, and the Workshop on Strategies, Prediction, Interaction, and Reasoning in Italy (AI4CC-IPS-RCRA-SPIRIT 2024), co-located with 23rd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2024), CEUR Workshop Proceedings, CEUR-WS.org, 2024.

[2] Intergovernmental Panel on Climate Change, Climate Change 2022 – Impacts, Adaptation and Vulnerability: Working Group II Contribution to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change, Technical Report, IPCC Cambridge University Press, 2023. doi:10.1017/9781009325844.

[3] United Nations (UN), The United Nations World Water Development Report 2023: Partnerships and cooperation for water, 2023.

[4] World Meteorological Organization (WMO), State of Global Water Resources, 2023.

[5] A. Wunsch, T. Liesch, G. Cinkus, N. Ravbar, Z. Chen, N. Mazzilli, H. Jourde, N. Goldscheider, Karst spring discharge modeling based on deep learning using spatially distributed input data, Hydrology and Earth System Sciences 26 (2022) 2405–2430. doi:10.5194/hess-26-2405-2022.

[6] H. Tao, M. M. Hameed, H. A. Marhoon, M. Zounemat-Kermani, S. Heddam, S. Kim, S. O. Sulaiman, M. L. Tan, Z. Sa'adi, A. D. Mehr, M. F. Allawi, S. Abba, J. M. Zain, M. W. Falah, M. Jamei, N. D. Bokde, M. Bayatvarkeshi, M. Al-Mukhtar, S. K. Bhagat, T. Tiyasha, K. M. Khedher, N. Al-Ansari, S. Shahid, Z. M. Yaseen, Groundwater level prediction using machine learning models: A comprehensive review, Neurocomputing 489 (2022) 271–308. doi:10.1016/j.neucom.2022.03.014.

[7] M. Salis, A. M. Atto, S. Ferraris, R. Meo, Time Distributed Deep Learning models for Purely Exogenous Forecasting. Application to Water Table Depth Prediction using Weather Image Time Series, 2024. doi:10.48550/arXiv.2409.13284. arXiv:2409.13284.

[8] S. Mohanty, M. K. Jha, A. Kumar, D. K. Panda, Comparative evaluation of numerical model and artificial neural network for simulating groundwater flow in Kathajodi–Surua Inter-basin of Odisha, India, Journal of Hydrology 495 (2013) 38–51. doi:10.1016/j.jhydrol.2013.04.041.

[9] A. Wunsch, T. Liesch, S. Broda, Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), Hydrology and Earth System Sciences 25 (2021) 1671–1687. doi:10.5194/hess-25-1671-2021.

[10] S. R. Clark, D. Pagendam, L. Ryan, Forecasting Multiple Groundwater Time Series with Local and Global Deep Learning Networks, International Journal of Environmental Research and Public Health 19 (2022) 5091. doi:10.3390/ijerph19095091.

[11] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, K.-R. Müller (Eds.), Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, volume 11700 of *Lecture Notes in Computer Science*, Springer International Publishing, Cham, 2019. doi:10.1007/978-3-030-28954-6.

[12] S. Letzgus, P. Wagner, J. Lederer, W. Samek, K.-R. Muller, G. Montavon, Toward Explainable Artificial Intelligence for Regression Models: A methodological perspective, IEEE Signal Processing Magazine 39 (2022) 40–58. doi:10.1109/MSP.2022.3153277.

[13] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, F. Herrera, Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence, Information Fusion 99 (2023) 101805. doi:10.1016/j.inffus.2023.101805.

[14] V. Petsiuk, A. Das, K. Saenko, Rise: Randomized input sampling for explanation of black-box models, in: British Machine Vision Conference (BMVC), 2018. URL: http://bmvc2018.org/contents/papers/1064.pdf.

[15] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, Association for Computing Machinery, New York, NY, USA, 2016, p. 1135–1144. URL: https://doi.org/10.1145/2939672.2939778. doi:10.1145/2939672.2939778.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, International Journal of Computer Vision 128 (2020) 336–359. doi:10.1007/s11263-019-01228-7.

[17] F. X. Gaya-Morey, J. M. Buades-Rubio, C. Manresa-Yee, Local Agnostic Video Explanations: A Study on the Applicability of Removal-Based Explanations to Video, 2024. doi:10.48550/arXiv.2401.11796. arXiv:2401.11796.

[18] S. Anderson, V. Radić, Evaluation and interpretation of convolutional long short-term memory networks for regional hydrological modelling, Hydrology and Earth System Sciences 26 (2022) 795–825. doi:10.5194/hess-26-795-2022.

[19] A. Stergiou, G. Kapidis, G. Kalliatakis, C. Chrysoulas, R. Veltkamp, R. Poppe, Saliency Tubes: Visual Explanations for Spatio-Temporal Convolutions, in: 2019 IEEE International Conference on Image Processing (ICIP), 2019, pp. 1830–1834. doi:10.1109/ICIP.2019.8803153. arXiv:1902.01078.

[20] Z. Li, W. Wang, Z. Li, Y. Huang, Y. Sato, Towards Visually Explaining Video Understanding

Networks With Perturbation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1120–1129.

[21] T. Uchiyama, N. Sogi, K. Niinuma, K. Fukui, Visually explaining 3D-CNN predictions for video classification with an adaptive occlusion sensitivity analysis, in: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2023, pp. 1513–1522. doi:10.1109/WACV56688.2023.00156.

[22] Q. Pan, W. Hu, J. Zhu, Series Saliency: Temporal Interpretation for Multivariate Time Series Forecasting, 2020. doi:10.48550/arXiv.2012.09324. arXiv:2012.09324.

[23] U. Schlegel, D. Oelke, D. A. Keim, M. El-Assady, An Empirical Study of Explainable AI Techniques on Deep Learning Models For Time Series Tasks, 2020. doi:10.48550/arXiv.2012.04344. arXiv:2012.04344.

[24] U. Schlegel, H. Arnout, M. El-Assady, D. Oelke, D. A. Keim, Towards A Rigorous Evaluation Of XAI Methods On Time Series, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 4197–4201. doi:10.1109/ICCVW.2019.00516.

[25] A. Theissler, F. Spinnato, U. Schlegel, R. Guidotti, Explainable AI for Time Series Classification: A Review, Taxonomy and Research Directions, IEEE Access 10 (2022) 100700–100724. doi:10.1109/ACCESS.2022.3207765.