INTERPRETABLE MACHINE LEARNING FOR EXTREME EVENTS DETECTION: AN APPLICATION TO DROUGHTS IN THE PO RIVER BASIN

Paolo Bonetti*
Politecnico di Milano

Matteo Giuliani
Politecnico di Milano

Veronica Cardigliano
Politecnico di Milano

Alberto Maria Metelli
Politecnico di Milano

Marcello Restelli
Politecnico di Milano

Andrea Castelletti
Politecnico di Milano

ABSTRACT

The increasing frequency and intensity of drought events—periods of significant decrease in water availability—are among the most alarming impacts of climate change. Monitoring and detecting these events is essential to mitigate their impact on our society. However, traditional drought indices often fail to accurately detect such impacts as they mostly focus on single precursors. In this study, we leverage machine learning algorithms to define a novel data-driven, impact-based drought index reproducing as target the Vegetation Health Index, a satellite signal that directly assesses the vegetation status. We first apply novel dimensionality reduction methods that allow for interpretable spatial aggregation of features related to precipitation, temperature, snow, and lakes. Then, we select the most informative and non-redundant features through filter feature selection. Finally, linear supervised learning methods are considered, given the small number of samples and the aim of preserving interpretability. The experimental setting focuses on ten sub-basins of the Po River basin, but the aim is to design a machine learning-based workflow applicable on a large scale.

1 INTRODUCTION

In recent years, climate change has gained increasing attention both in the media and within the scientific community (Parmesan et al., 2022). An alarming consequence of climate change and global warming is the rising frequency and intensity of droughts (Spinoni et al., 2016), defined as periods of aridity and scarcity of water compared to normal conditions (Van Loon & Van Lanen, 2012), which can have severe economic and ecological effects, impacting agriculture, water resources, tourism, ecosystems, especially in regions lacking effective mitigation and drought management plans (Dai, 2011). In this scenario, the identification of drought-prone conditions and early detection are crucial to mitigate their consequences.

Traditional standardized drought indices (e.g., SPI (Guttman, 1999), SPEI (Vicente-Serrano et al., 2010)) are widely used to monitor drought conditions. However, they often fail to detect drought impacts as they focus on specific drivers (e.g., precipitation and evapotranspiration) without accounting for their complex interactions, that eventually generate the drought's impacts. Ad hoc index formulations have also been designed to identify drought conditions in specific basins (e.g., Estrela & Vargas (2012)). However, these indices are basin-specific, with no capability to generalize and no automatization of the process, requiring years of expert refinements. As an alternative, indices derived from satellite data are available (e.g., NDVI (Pettorelli, 2013), VHI (Bento et al., 2018)) and allow the direct observation of vegetation status in terms of colors, making them more suitable proxies of its condition. Yet, these indices can only be observed in real-time, preventing their prediction weeks to months in advance (Meehl et al., 2021) to timely prompt anticipatory operations.

*paolo.bonetti@polimi.it

Recent works applied Machine Learning to automatically derive data-driven drought indices, exploiting classical indices as proxies of drought conditions (e.g., (Zaniolo et al., 2018; Feng et al., 2019; Dikshit et al., 2021)). Building on these works, we apply a similar data-driven workflow in the Po River basin (Italy), which is the largest Italian catchment and the most populated area of the country. This region plays a central role in the national economy, and it is one of the largest agricultural areas in Europe. In this context, we consider the remotely sensed Vegetation Health Index (Vogt et al., 2000, VHI) as a proxy of the state of vegetation impacted by drought conditions. The novel contribution of this work resides on the interpretability of each step of the ML workflow proposed. Indeed, this work is the first empirical validation of recent dimensionality reduction algorithms (theoretically presented in (Bonetti et al., 2024; 2023), designed to obtain interpretable spatially aggregated features. Then, filter feature selection through conditional mutual information (CMI) is used to select the relevant and non-redundant features as candidate drivers. Finally, linear supervised learning models are trained, also considering a multi-task setting, in order to define a new data-driven, impact-based drought index (i.e., the resulting model) reproducing the state of vegetation impacted by the drought. The results show the effectiveness of the proposed workflow, whose main contribution is not only in terms of regression performances but also in terms of index interpretability, which is essential to bridge the gap between advanced ML methods (often considered as black-box algorithms) and the physical meaning of the results.

2 DATA

The VHI is a satellite index available in eight-day aggregations and 231m spatial resolution (Zellner & Castelli, 2022). We worked on ten hydrological sub-basins (Figure 1a) of the Po River basin (Lehner & Grill, 2013a), where urbanized regions were excluded to focus on cultivable areas only[1]. The target of each sub-basin is, therefore, the eight-day average VHI value of cultivable areas, representing a proxy of the state of vegetation.

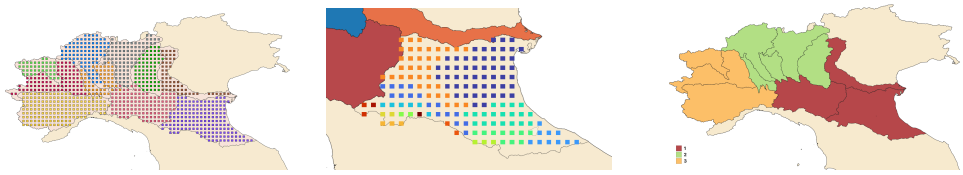
The initial set of candidate drivers, i.e., input features, are meteorological features. Daily temperature and cumulative precipitation (Cornes et al., 2018) have been considered since they are the core variables of traditional drought indices. Then, daily snow depth cover (Hersbach et al., 2023) for sub-basins that include mountainous areas have been added. Snow depth responds to temperature fluctuations and can impact water availability through snowmelt, potentially mitigating or delaying drought conditions. Finally, lake-related features, i.e., water level, inflow, and release, have also been included since they may be indicators of prolonged water scarcity. Specifically, data from the four largest lakes in the region - Lake Como, Lake Iseo, Lake Maggiore, and Lake Lugano - were considered.

Starting from daily features, they have been aggregated into 8-day averages, matching the temporal resolution of the target. Given the possible long-term impact and delayed effect, averaged variables over $0, 1, 4, 8, 12, 16, 24$ weeks have been considered. Then, to take into account cyclo-stationarity and remove seasonality, the average value (based on training set) of each feature associated with a particular week has been removed to weekly measurement.

Since snow and lake features influence only a subset of the considered sub-basins, the main focus of the analysis is on precipitation and temperature data. For these variables, we considered the original grid points at different locations as individual features. In the considered domain, there are 991 distinct latitude-longitude grid points across the 10 sub-basins. Considering their 14 temporal aggregations, we have a total of 14×991 initial features. As a first novel contribution, we processed this data using a novel data-driven feature aggregation algorithm combined with a filter feature selection (see Section 3).

To validate our results, we also run a benchmark experiment that replaces the dimensionality reduction step and directly calculates mean feature values for each variable in each sub-basin. In this case, we obtain 14×10 candidate drivers, which are the mean value of daily temperature and precipitation anomalies over the area of each sub-basin, temporally averaged over 8-day intervals and over the preceding 0, 1, 4, 8, 12, 16, and 24 blocks of 8 days.

[1]Following the masking available at: Land Cover Map 2021



(a) Study area and coordinates of temperature and precipitation. The same pixel's color refers to the same sub-basin.

(b) Example of application of NonLinCFA on 1-week precipitation. Each color refers to an aggregated feature.

(c) Representation of the three clusters of basins considered for the multi-task setting.

Figure 1: Illustrative examples: sub-basins, NonLinCFA aggregation, target clusters.

Given the availability of VHI data, the study period spans from January 2001 to December 2019, for a total of 867 samples. The dataset has been divided into training (2001-2009, 411 samples), validation (2010-2014, 228 samples), and test set (2015-2019, 228 samples).

3 Extraction of Interpretable Relevant Features

Dimensionality Reduction This work builds upon three recent dimensionality reduction approaches, LinCFA, NonLinCFA, and GenLinCFA (Bonetti et al., 2024; 2023)), designed to identify groups of features that are convenient, in terms of MSE, to aggregate with their mean. In this way, the interpretability of the reduced features is preserved, as they are averages of a subset of features. Iteratively, these algorithms compare pairs of features and aggregate them with the mean, depending on a threshold based on correlation (we refer the interested reader to the original papers for further details). *Linear Correlated Features Aggregation* (LinCFA) assumes linear relationships between the features and the target. *Non-Linear Correlated Features Aggregation* (NonLinCFA) is an extension of LinCFA designed assuming that the target is a nonlinear function of the features, with additive Gaussian noise, which is more realistic. Finally, *Generalized Linear Correlated Features Aggregation* (GenLinCFA) is a further extension that assumes that the distribution of the target belongs to the canonical exponential family, making the algorithm applicable in classification settings.

We apply these algorithms to the temperature and precipitation data described in Section 2, where 991 values of each feature at different locations are available. Specific heuristics have also been introduced in our implementation. In particular, only features belonging to the same sub-basin, with neighboring locations, and representing the same variable (e.g., temperature) over the same timescale (e.g., 24 weeks) can be aggregated together. Additionally, an internal aggregation order prioritizing pairs of features with the highest correlation has been added to avoid variability due to random ordering. In conclusion, the reduced features produced are averages across neighboring grid points within the same region of the same variable, where the most promising features (the most correlated ones) are first checked for aggregation. Figure 1b visually shows the aggregations identified by the NonLinCFA algorithm, considering the 1-week aggregated cumulative precipitation feature in a specific sub-basin. It reveals the identification of different sub-regions and the extraction of a total of 17 features, starting from 172 original ones. A further classification analysis applying GenLinCFA has been performed and it is described in the appendix.

Feature Selection Feature selection based on CMI has subsequently been applied to identify relevant and non-redundant features, i.e., drivers of drought conditions. The supervised filter algorithm considered selects features in a forward or backward manner, it is model independent, and provides theoretical guarantees on the loss of information (Beraha et al., 2019). We apply forward CMI feature selection to each sub-basin, using the corresponding VHI as target and the features obtained from dimensionality reduction as inputs. Additionally, in the benchmark experiment, feature selection is applied directly to averaged features. In both cases, we identify the five most relevant features for each sub-basin, selected according to their performance on the validation set. As an additional benchmark, we explored a forward wrapper feature selection method. Our results (see Appendix) show that the CMI-based approach performs comparably or even better than the wrapper method, with the advantage of being model-independent. This preserves the interpretability of the selected

	Adda	Dora	Emiliani1	Emiliani2	Garda_Mincio	Lambro_Olona	Oglio_Iseo	Piemonte_Sud	Piemonte_Nord	Ticino
R^2 NonLinCFA (MTL)	0.1 (**0.21**)	-0.2 (-0.13)	0.29 (**0.33**)	0.24 (**0.29**)	0.17 (**0.23**)	**0.26** (0.24)	0.2 (**0.22**)	0.12 (-0.09)	0.06 (0.03)	0.16 (**0.23**)
R^2 Averages	0.16	-0.15	0.32	0.22	0.19	0.21	0.20	**0.15**	**0.12**	0.17
R^2 SMA	0.01	-0.11	0.03	0.02	0.05	0.08	0.05	0.12	0.02	0.06

Table 1: Test scores in terms of coefficient of determination (R^2) of linear regressions in single and multi-task (in parenthesis) for the ten sub-basins. NonLinCFA-based results are compared with a benchmark referring to features averaged over the whole sub-basin (second row), and a baseline based on an agricultural drought index (SMA, third row).

candidate drivers since it would not be physically meaningful to select different meteorological variables as drivers of the same drought condition depending on the ML model considered.

4 IDENTIFICATION OF DATA-DRIVEN, IMPACT-BASED DROUGHT INDICES

The relevant features selected by the approach described in the previous section are used as inputs for simple supervised learning models to reconstruct the target VHI in each sub-basin. A multi-task variation was also explored, grouping the ten sub-basins into three clusters based on VHI correlations and DBSCAN clustering (Figure 1c). The clusters are also meaningful in terms of proximity and hydrological characteristics. In this case, a linear model was trained for each cluster, considering input features from all its basins and a bias term through one-hot encoding, which associates each sample to the related sub-basin. This multi-task approach provides benefits such as increased sample size and reduced number of models.

Table 1 illustrates the results, in terms of coefficient of determination (R^2), for single-task and multi-task linear regressions (first row), compared with two baselines: a linear regression that considers the average of each feature as input (second row) and a linear regression that considers as input Soil Moisture Anomaly (Bergman et al., 1988, SMA). This way, the proposed workflow is compared with a model that does not consider NonLinCFA algorithm and another one based on an agricultural drought index. The results highlight improvements with the proposed workflow, especially in the multi-task setting. In most cases, the test R^2 is above 0.2, showing that we are extracting a significantly positive amount of information, although we do not achieve test scores higher than 0.35, which may depend on the small number of samples, the difficulty of the task (e.g., the score associated with SMA is always close to 0), and the fact that we are testing on 5 years in the future w.r.t. the training set.

More complex ML models, such as feed-forward neural networks and random forests, have also been considered. However, we did not find significant improvements, which may depend on the limited amount of data samples. Additionally, linear models have been preferred to comply with the goal of preserving the interpretability of the results. Also, the addition of snow depth and lake-related features does not enhance the accuracy of the models, suggesting that their contribution is not significantly complementary to precipitation and temperature. We also considered a classification setting where the target represents the presence of drought conditions. Accuracy scores, considering logistic regression, are reported in the appendix.

5 CONCLUSIONS

This work addresses drought detection through the reconstruction of a data-driven index, considering meteorological features as input variables and a satellite signal measuring the state of the vegetation as target. The implemented ML pipeline includes a first empirical validation of recent dimensionality reduction algorithms, combined with filter feature selection and a linear multi-task learning approach. The test scores obtained are overall satisfactory, given the complexity of the problem and the small number of samples, although they may be improved in future work. The main contribution of this work is not only in the direction of obtaining an accurate prediction, but to the preservation of interpretability. The resulting indices are linear combinations of drivers selected among a set of features that have been extracted as averages of subsets of original meteorological variables at neighbouring locations. Therefore, these data-driven indices are physically interpretable, and they can be therefore analyzed by domain experts. Future work will focus on testing a broader application

of the proposed pipeline at the European scale to assess its capability to automatize the data-driven extraction of drought indices on a larger domain.

ACKNOWLEDGEMENTS

This work has been supported by the CLINT research project funded by the H2020 Programme of the European Union under Grant Agreement No 101003876. This paper is supported by PNRR-PE-AI FAIR project funded by the NextGeneration EU program.

REFERENCES

Virgílio A Bento, Célia M Gouveia, Carlos C DaCamara, and Isabel F Trigo. A climatological assessment of drought impact on vegetation health index. *Agricultural and forest meteorology*, 259:286–295, 2018.

Mario Beraha, Alberto Maria Metelli, Matteo Papini, Andrea Tirinzoni, and Marcello Restelli. Feature selection via mutual information: New theoretical insights. In *2019 international joint conference on neural networks (IJCNN)*, pp. 1–9. IEEE, 2019.

KH Bergman, P Sabol, and D Miskus. Experimental indices for monitoring global drought conditions. In *Proceedings of the 13th annual climate diagnostics workshop, Cambridge, MA, USA*, volume 31, pp. 190–197, 1988.

Paolo Bonetti, Alberto Maria Metelli, and Marcello Restelli. Nonlinear feature aggregation: Two algorithms driven by theory, 2023.

Paolo Bonetti, Alberto Maria Metelli, and Marcello Restelli. Interpretable linear dimensionality reduction based on bias-variance analysis. *Data Mining and Knowledge Discovery*, pp. 1–69, 2024.

Richard C. Cornes, Gerard van der Schrier, Else J. M. van den Besselaar, and Philip D. Jones. An ensemble version of the e-obs temperature and precipitation data sets. *Journal of Geophysical Research: Atmospheres*, 123(17):9391–9409, 2018. doi: https://doi.org/10.1029/2017JD028200. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2017JD028200.

Aiguo Dai. Drought under global warming: a review. *Wiley Interdisciplinary Reviews: Climate Change*, 2(1):45–65, 2011.

David A Dickey and Wayne A Fuller. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association*, 74(366a):427–431, 1979.

Abhirup Dikshit, Biswajeet Pradhan, and Alfredo Huete. An improved spei drought forecasting approach using the long short-term memory neural network. *Journal of environmental management*, 283:111979, 2021.

Teodoro Estrela and Elisa Vargas. Drought management plans in the european union. the case of spain. *Water resources management*, 26(6):1537–1553, 2012.

Puyu Feng, Bin Wang, De Li Liu, and Qiang Yu. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in south-eastern australia. *Agricultural Systems*, 173:303–316, 2019.

Nathaniel B Guttman. Accepting the standardized precipitation index: a calculation algorithm 1. *JAWRA Journal of the American Water Resources Association*, 35(2):311–322, 1999.

H. Hersbach, B. Bell, P. Berrisford, G. Biavati, A. Horányi, J. Muñoz Sabater, J. Nicolas, C. Peubey, R. Radu, I. Rozum, D. Schepers, A. Simmons, C. Soci, D. Dee, and J-N. Thépaut. Era5 hourly data on single levels from 1940 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS), 2023.

Denis Kwiatkowski, Peter C.B. Phillips, Peter Schmidt, and Yongcheol Shin. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, 54(1):159–178, 1992. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(92)90104-Y. URL https://www.sciencedirect.com/science/article/pii/030440769290104Y.

Bernhard Lehner and Günther Grill. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15): 2171–2186, 2013a.

Bernhard Lehner and Günther Grill. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15): 2171–2186, 2013b. doi: https://doi.org/10.1002/hyp.9740. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hyp.9740.

Greta Ljung and G. Box. On a measure of lack of fit in time series models. *Biometrika*, 65, 08 1978. doi: 10.1093/biomet/65.2.297.

Gerald A. Meehl, Jadwiga H. Richter, Haiyan Teng, Antonietta Capotondi, Kim Cobb, Francisco Doblas-Reyes, Markus G. Donat, Matthew H. England, John C. Fyfe, Weiqing Han, Hyemi Kim, Ben P. Kirtman, Yochanan Kushnir, Nicole S. Lovenduski, Michael E. Mann, William J. Merryfield, Veronica Nieves, Kathy Pegion, Nan Rosenbloom, Sara C. Sanchez, Adam A. Scaife, Doug Smith, Aneesh C. Subramanian, Lantao Sun, Diane Thompson, Caroline C. Ummenhofer, and Shang-Ping Xie. Initialized Earth System prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, 2(5):340–357, 2021. ISSN 2662-138X. doi: 10.1038/s43017-021-00155-x. URL https://doi.org/10.1038/s43017-021-00155-x.

Camille Parmesan, Mike D Morecroft, and Yongyut Trisurat. *Climate change 2022: Impacts, adaptation and vulnerability*. PhD thesis, GIEC, 2022.

Nathalie Pettorelli. *The normalized difference vegetation index*. Oxford University Press, USA, 2013.

QGIS.org. Qgis software. http://www.qgis.org, 2022. QGIS Geographic Information System. QGIS Association.

Jonathan Spinoni, Gustavo Naumann, Jürgen Vogt, and Paulo Barbosa. Meteorological droughts in europe: events and impacts-past trends and future projections. 2016.

Anne F Van Loon and Henny AJ Van Lanen. A process-based typology of hydrological drought. *Hydrology and Earth System Sciences*, 16(7):1915–1946, 2012.

Sergio M Vicente-Serrano, Santiago Beguería, and Juan I López-Moreno. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *Journal of climate*, 23(7):1696–1718, 2010.

J. V. Vogt, S. Niemeyer, F. Somma, I. Beaudin, and A. A. Viau. *Drought Monitoring from Space*, pp. 167–183. Springer Netherlands, Dordrecht, 2000. ISBN 978-94-015-9472-1. doi: 10.1007/978-94-015-9472-1_13. URL https://doi.org/10.1007/978-94-015-9472-1_13.

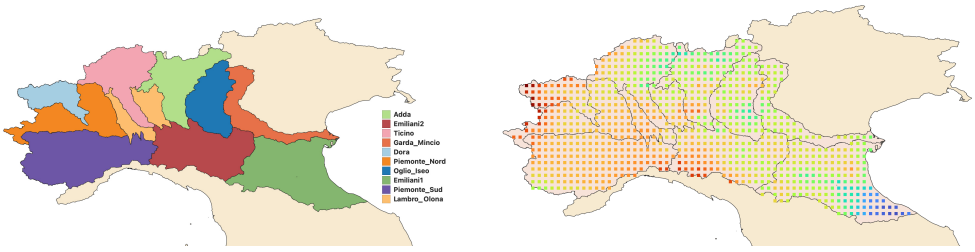
Marta Zaniolo, Matteo Giuliani, Andrea Francesco Castelletti, and Manuel Pulido-Velazquez. Automatic design of basin-specific drought indexes for highly regulated water systems. *Hydrology and Earth System Sciences*, 22(4):2409–2424, 2018.

P. Zellner and M. Castelli. *Vegetation Health Index - 231 m 8 days (Version 1_0) [Data set]*. Eurac Research, 2022.

A DATA

This appendix section provides additional details on the data considered for this work.

The focus of this study is the area of the Po River basin, as illustrated in Figure 2a. Within this region, we have identified ten primary hydrological sub-basins whose labels are in the figure legend. These sub-basins are established based on the partition of hydrological levels 7 and 8, among those detailed in (Lehner & Grill, 2013b).



(a) Study area considered, divided into ten hydrological sub-basins whose names and relative colour appear in the legend. Shapefiles created using QGIS software (QGIS.org, 2022).

(b) Plot of temperature feature values related to a sample week without temporal aggregations.

Figure 2: Study area and example of temperature values.

The chosen target signal for this drought detection case study is the VHI. Firstly, we examined the autocorrelation and stationarity properties of the VHI time series. Our analysis confirmed the stationarity of the signal, as it passed both the Augmented Dickey-Fuller (ADF) (Dickey & Fuller, 1979) and Kwiatkowski-Phillips-Schmidt-Shin (KPSS) (Kwiatkowski et al., 1992) tests. Furthermore, the VHI time series exhibited an autoregressive behaviour characterized by various orders of autocorrelation. This autoregressive nature was supported by the accuracy of autoregressive model predictions, which were compared against the actual test signal (Figure 3), and by the results of the Ljung-Box test (Ljung & Box, 1978) applied to the residuals. The absence of correlation among residuals indicates that they behave like white noise.

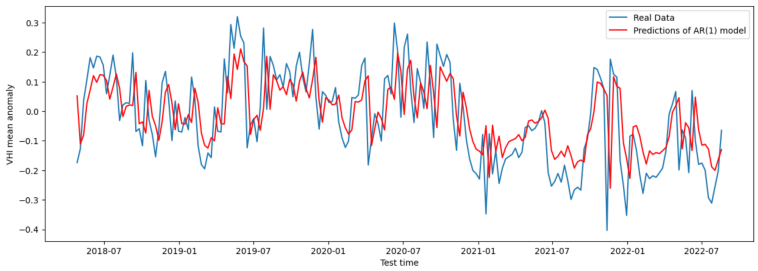


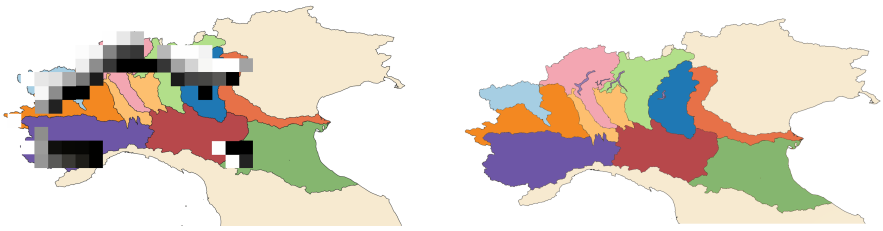
Figure 3: VHI target predictions of AR(1) model with respect to the true signal.

However, the primary objective of this work is not to reconstruct the signal precisely but rather to leverage it for the data-driven detection of features that influence its fluctuations. As a result, we focused on building models for signal reconstruction based on meteorological data.

Our main dataset comprises cumulative precipitation and daily temperature features, available in a gridded format and covering the entire region of interest. As an example, Figure 2b displays temperature feature values without temporal aggregations on a specific week. These values undergo a cyclostationary transformation during preprocessing. For this reason, red points indicate a higher anomaly compared to the mean for that specific week of the year and coordinate, whereas blue points indicate a lower anomaly, meaning values close to the mean for that point during that time of year. From this plot, we can observe how feature values exhibit significant variations even within the same sub-basin. The benchmark experiment, introduced in Section 2, does not account for these variations,

as it computes a single mean feature per sub-basin by averaging all values. Dimensionality reduction methods, instead, enable the identification of similar areas and facilitate the selection of the most informative ones with respect to the VHI target of a specific sub-basin.

In addition to cumulative precipitation and daily temperature, we incorporated data related to snow depth and lakes in our feature set. Specifically, snow data are available in a grid format and are considered only at points where snowfall occurs at least sporadically, as illustrated in Figure 4a. Regarding the lakes, we utilized information related to water level, inflow and release. Figure 4b displays the locations of the considered lakes.



(a) Plot of snow depth values in a sample date from (Hersbach et al., 2023), where the darker squares represent lower values of the variable, and snow-free points have been excluded.

(b) Geographical location of the lakes of interest, colored in purple, within the sub-basins (the ten colored regions). The surrounding area represents the rest of the North Italian territory.

Figure 4: Geographical identification of snow depth values and lakes.

B METHODS AND RESULTS

This second appendix section provides further details and results of the ML methodology followed to address the detection task.

As described in the main paper, after the extraction and analysis of the target and feature variables, we followed the workflow visually summarized in the block scheme in Figure 5. The initial step involves applying the discussed dimensionality reduction techniques to spatio-temporal features. Subsequently, feature selection using the CMI approach is performed on the spatially aggregated features obtained in the previous step. These selected features are then employed in the final phase, where supervised learning models are trained to reconstruct the VHI target. For the regression task,

Figure 5: Block scheme of the workflow of the conducted experiments.

the models were trained with features derived from the NonLinCFA dimensionality reduction method, considering a continuous target. NonLinCFA spatially aggregates variable values across all basin points, employing a theoretical threshold to decide which points to merge together and replace by their mean. Following this, feature selection methods are applied to identify the most informative and non-redundant inputs among the spatial aggregations generated for each temporal aggregation of each variable. A practical example of the results for a sub-basin is provided in Figure 6. Here, each row refers to one of the top 5 features selected by the forward CMI feature selection algorithm. The feature names are reported above each plot, with *tg* denoting daily temperature variables, *rr* referring to cumulative precipitation ones and *nw* indicating that each value of that feature is the average of that variable *n* weeks before. In particular, the left column displays all spatial aggregations generated

using the NonLinCFA algorithm for the respective feature. These aggregations represent new features derived from averaging the variable values across all the data points included in each aggregation. The right column, instead, shows only the aggregation selected as one of the most relevant during the feature selection phase. We can observe how NonLinCFA identifies a variety of different aggregations based on the similarity between points, which can be easily visualized and analyzed.

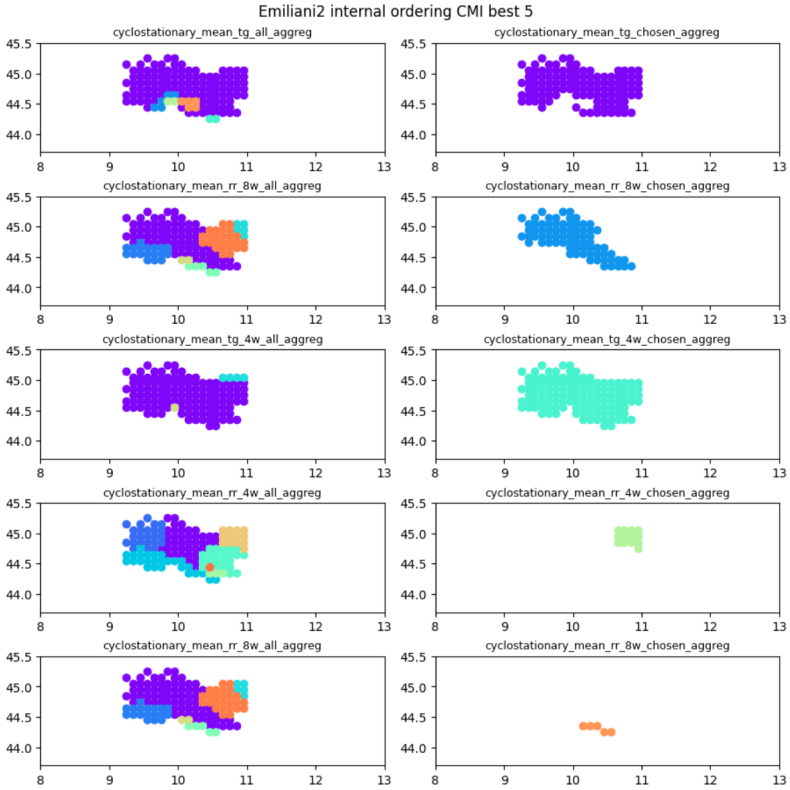


Figure 6: Aggregations on *Emiliani2* basin using the NonLinCFA algorithm and selecting the first 5 features from forward CMI feature selection.

The models were trained both in single-task and multi-task settings. As illustrated in Figure 1c, three clusters of basins have been identified based on their similarity. For multi-task training, all the chosen features from each basin within a cluster are employed as inputs for the model. Additionally, we incorporate a one-hot encoding to provide the model with information about which sub-basin each sample belongs to. An example of the obtained dataframe is displayed in Figure 7.

Table 2 is an extended version of Table 1, where also results from wrapper feature selection methods have been included.

Input features derived from the wrapper feature selection method slightly improved performance in single-task contexts for certain sub-basins, such as *Emiliani2* and *Garda Mincio*. However, the most favourable results in the multi-task setting are usually achieved by employing features selected through the CMI-based feature selection method. This is exemplified by the performance leap observed for the *Emiliani2* sub-basin in the single-task context, which is compensated in the multi-task scenario, where the CMI score surpasses the one obtained with the wrapper method. This empirically validates the robust generalization capability of features chosen based on CMI with respect to the ones obtained through the wrapper feature selection method, which is more computationally expensive and related to the specific trained model.

The last column of Table 2 refers to the benchmark results obtained with features averaged across the entire sub-basin area. Notably, the application of NonLinCFA results in improvements, particularly pronounced in the multi-task setting. Some sub-basins such as *Dora*, *Piemonte Sud*, and *Piemonte*

Garda_Mincio_cyclostationary_mean_tg_1w_1	Garda_Mincio_cyclostationary_mean_rr_4w_3	Emiliani1	Emiliani2	Garda_Mincio
-1.218678	1.758769	1	0	0
-0.434251	1.121893	1	0	0
-1.019494	0.598275	1	0	0
-1.130695	0.310672	1	0	0
-0.942006	0.314132	1	0	0
...	...	...	...	...
1.683961	0.030398	0	0	1
1.024164	0.377049	0	0	1
0.349305	0.643723	0	0	1
-1.418043	-0.102080	0	0	1
-1.763499	0.187242	0	0	1

Figure 7: Features dataframe structure in multi-task settings. The boolean columns indicate the basin to which each sample belongs.

Nord proved to be more challenging to address, possibly due to the specific mountainous precipitation configurations and to the small number of cultivable areas available.

	NonLinCFA wrapper best 5 (MTL)	NonLinCFA CMI best 5 (MTL)	CMI best 5 on cont. target
Adda	0.09 (0.08)	0.1 (0.21)	0.16
Dora	-0.07 (-0.14)	-0.2 (-0.13)	-0.15
Emiliani1	0.24 (0.27)	0.29 (0.33)	0.32
Emiliani2	0.27 (0.22)	0.24 (0.29)	0.22
Garda Mincio	0.21 (0.23)	0.17 (0.23)	0.19
Lambro Olona	0.2 (-0.07)	0.26 (0.24)	0.21
Oglio Iseo	0.28 (0.08)	0.2 (0.22)	0.2
Piemonte Sud	0.15 (0.0)	0.12 (-0.09)	0.15
Piemonte Nord	0.04 (0.05)	0.06 (0.03)	0.12
Ticino	0.22 (0.07)	0.16 (0.23)	0.17

Table 2: R^2 scores in test of Linear Regression model in single-task and multi-task (in parenthesis) for all the ten sub-basins. NonLinCFA results have to be compared with the benchmark in the last column, where CMI has been applied to features averaged across the whole sub-basin.

The outcomes that were detailed in the single-task and multi-task regression scenarios are also presented in the context of binary classification. In this case, we evaluated the model performance in distinguishing between *Good* and *Bad* classes, respectively, referring to unfavourable and favourable conditions for drought appearance. The binary target was derived by categorizing the VHI signal into these two classes, and we explored various thresholds to perform this division in the most effective and significant way. Table 3 provides an overview of the accuracy scores obtained through the logistic regression model and the GenLinCFA dimensionality reduction method, leveraging daily temperature and cumulative precipitation features. While the scores in this context exhibit a coarser granularity, and the deviation from the benchmark (achieved by averaging features across entire sub-basins) and from the wrapper feature selection results is less pronounced, the performance remains consistently comparable. In addition, the results are achieved using faster, model-independent, and interpretable algorithms, which still make it competitive.

	GenLinCFA wrapper best 5 (MTL)	GenLinCFA CMI best 5 (MTL)	CMI best 5 on discrete target (MTL)
Adda	0.65 (0.68)	0.67 (0.65)	0.62
Dora	0.56 (0.62)	0.55 (0.58)	0.5
Emiliani1	0.69 (0.73)	0.7 (0.69)	0.72
Emiliani2	0.76 (0.7)	0.58 (0.73)	0.69
Garda Mincio	0.76 (0.7)	0.73 (0.73)	0.71
Lambro Olona	0.67 (0.66)	0.69 (0.65)	0.68
Oglio Iseo	0.68 (0.68)	0.71 (0.64)	0.61
Piemonte Sud	0.68 (0.69)	0.63 (0.63)	0.63
Piemonte Nord	0.68 (0.68)	0.61 (0.63)	0.67
Ticino	0.69 (0.68)	0.65 (0.67)	0.68

Table 3: Accuracy scores in test of Logistic Regression model in single-task and multi-task (in parenthesis) for all the ten sub-basins. GenLinCFA results have to be compared with the benchmark in the last column, where CMI has been applied to features averaged across the whole sub-basin.

The identical experiments were replicated while incorporating the snow depth and lake-related features, and the results are presented in Table 4. For this analysis, we focused only on sub-basins where these additional features were available, excluding the *Emiliani1 - Emiliani2 - Garda Mincio* cluster, which is predominantly flat and far from the considered lakes. The *Piemonte Nord - Piemonte Sud - Dora* cluster was also omitted since it does not contain the designated lakes.
When included, snow depth and lake behaviour features were identified as informative by the feature selection methods. However, their inclusion did not lead to a significant enhancement in the R^2 scores for regression. This observation indicates that while these features do provide valuable information, their contribution is not significantly complementary to that of precipitation and temperature data. Consequently, incorporating these features did not result in a noteworthy improvement in prediction performance, although we were able to extract useful information from them.

	NonLinCFA wrapper best 5 (MTL)	NonLinCFA CMI best 5 (MTL)	CMI best 5 on cont. target
Adda	0.09 (0.18)	0.1 (0.26)	0.16
Lambro Olona	0.2 (0.09)	0.26 (0.14)	0.21
Oglio Iseo	0.29 (0.19)	0.18 (0.25)	0.2
Ticino	0.14 (0.16)	0.16 (0.21)	0.17

Table 4: R^2 scores in test of Linear Regression model in single-task and multi-task (in parenthesis) for all the ten sub-basins. In this case, the considered features are not only daily temperature and cumulative precipitation, but also lakes and snow-related features.