POLITECNICO DI TORINO
Repository ISTITUZIONALE

HydroChronos: Forecasting Decades of Surface Water Change

Original
HydroChronos: Forecasting Decades of Surface Water Change / Rege Cambrin, D., Poeta, E., Pastor, E., Corley, I., Cerquitelli, T., Baralis, E., Garza, P.. - (2025), pp. 265-276. (33rd ACM International Conference on Advances in Geographic Information Systems Minneapolis (USA) November 3 - 6, 2025) [10.1145/3748636.3762732].

Availability:
This version is available at: 11583/3003218 since: 2026-02-19T19:59:23Z

Publisher:
ACM

Published
DOI:10.1145/3748636.3762732

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

30 October 2026

HydroChronos: Forecasting Decades of Surface Water Change

Daniele Rege Cambrin*
Politecnico di Torino, Italy
daniele.regecambrin@polito.it

Eleonora Poeta
Politecnico di Torino, Italy
eleonora.poeta@polito.it

Eliana Pastor
Politecnico di Torino, Italy
eliana.pastor@polito.it

Isaac Corley
Wherobots, USA
isaac@wherobots.com

Tania Cerquitelli
Politecnico di Torino, Italy
tania.cerquitelli@polito.it

Elena Baralis
Politecnico di Torino, Italy
elena.baralis@polito.it

Paolo Garza
Politecnico di Torino, Italy
paolo.garza@polito.it

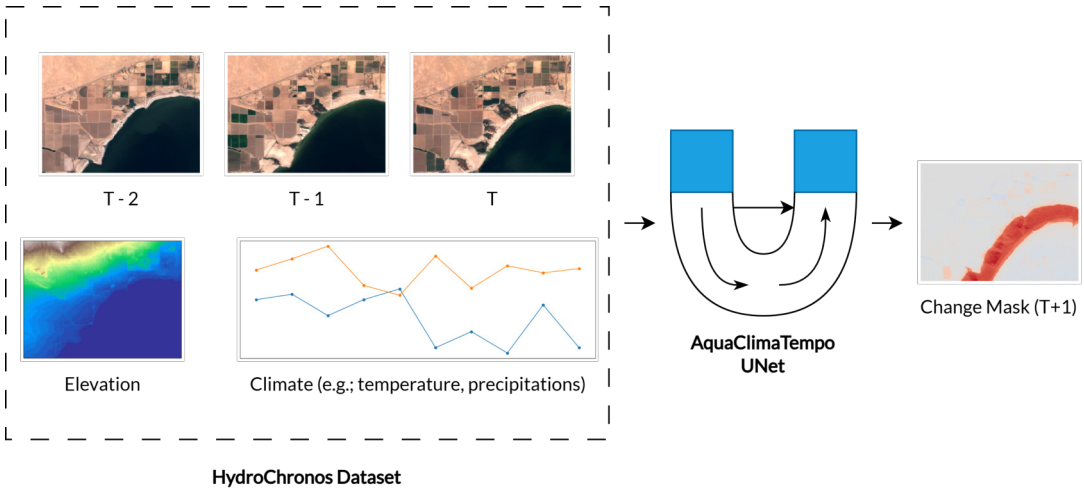


Figure 1: Forecasting future water landscapes: The HydroChronos dataset (left) provides rich multi-modal inputs including satellite image series, elevation, and climate data. Our *AquaClimaTempo UNet* (center) processes this information to predict future surface water dynamics, generating a *change mask* for the forecasted water dynamics for the next timestep (T+1) (right).

Abstract

Forecasting surface water dynamics is crucial for water resource management and climate change adaptation. However, the field lacks comprehensive datasets and standardized benchmarks. In this paper, we introduce HydroChronos, a large-scale, multi-modal spatiotemporal dataset for surface water dynamics forecasting designed to address this gap. We couple the dataset with three forecasting tasks. The dataset includes over three decades of aligned Landsat 5 and Sentinel-2 imagery, climate data, and Digital Elevation Models for diverse lakes and rivers across Europe, North America, and South America. We also propose AquaClimaTempo UNet, a novel spatiotemporal architecture with a dedicated climate data branch, as a strong benchmark baseline. Our model significantly outperforms a Persistence baseline for forecasting future water dynamics by +14% and +11% F1 across change detection and direction of change classification tasks, and by +0.1 MAE on the magnitude of change regression. Finally, we conduct an Explainable AI analysis to identify the key climate variables and input channels that influence surface water change, providing insights to inform and guide future modeling efforts.

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Applied computing** → **Environmental sciences**; • **Information systems** → **Geographic information systems**.

Keywords

Spatiotemporal Forecasting, Surface Water Dynamics, Remote Sensing, Explainable AI, Multi-modal Data Fusion

This work is licensed under a Creative Commons Attribution 4.0 International License.
SIGSPATIAL '25, Minneapolis, MN, USA
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2086-4/2025/11
https://doi.org/10.1145/3748636.3762732

ACM Reference Format:
Daniele Rege Cambrin, Eleonora Poeta, Eliana Pastor, Isaac Corley, Tania Cerquitelli, Elena Baralis, and Paolo Garza. 2025. HydroChronos: Forecasting Decades of Surface Water Change. In *The 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25), November 3–6, 2025, Minneapolis, MN, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3748636.3762732

1 Introduction

The escalating global challenges of water scarcity, climate change, and their profound impacts on ecosystems and human societies underscore the critical importance of understanding and forecasting surface water dynamics [34, 40]. Effective water resource management for agriculture, energy, and consumption relies on predicting future water availability. As climate change intensifies droughts and floods, robust forecasting models are indispensable for adaptation, enabling interventions and improving resilience [12, 18]. Spatio-temporal variability of surface water requires advanced analytics to capture complex hydrological processes and environmental responses [39]. Despite the pressing need for accurate and long-term surface water predictions, the research landscape is hampered by significant limitations. A primary obstacle is the lack of comprehensive, large-scale datasets specifically curated for forecasting tasks. Satellite observations are often fragmented or not integrated with crucial climate and topographic data. Consequently, there is also a scarcity of well-defined predictive tasks that leverage these multi-modal data sources to forecast surface water dynamics.

To fill the gap, this paper introduces HydroChronos, a novel, large-scale, multi-modal spatio-temporal dataset specifically designed to foster research in surface water forecasting. HydroChronos is characterized by an extensive temporal coverage, encompassing over three decades of Landsat 5 and Sentinel-2 satellite imagery. This imagery is integrated with corresponding climate variables (e.g., precipitation and temperature) and a Digital Elevation Model (DEM) for a diverse set of lake and river systems across Europe, the United States, and Brazil. Using this dataset, we define three standardized predictive tasks for surface water dynamics forecasting from satellite imagery (optionally with climate data): binary change detection, direction of change classification, and magnitude of change regression.

Building upon the HydroChronos dataset and the defined forecasting tasks, we propose a robust baseline model: AquaClimaTempo UNet (ACTU). This model, based on UNet [32] and ConvLSTM [36], features a climate data branch to learn interactions between historical water dynamics and climatic drivers. Our experimental results demonstrate that this model significantly outperforms the commonly used persistence baseline in forecasting future water dynamics. Furthermore, to foster transparency and understanding, we conduct an Explainable AI (XAI) analysis. This analysis offers insights into model decisions, identifies key drivers of surface water changes, and guides future research.

The contributions of this paper (Figure 1) can be summarized as follows:

- We introduce HydroChronos, the first dataset tailored for water dynamics forecasting, including remote-sensed images, climate variables, and DEM.
- We introduce three tasks of surface water dynamics forecasting, offering a benchmark for future research in spatio-temporal predictive modeling
- We introduce ACTU as a baseline model with the possibility to integrate climate variables and DEM
- We performed an XAI analysis on our models to guide future research and understand the influence of various factors

The code and the dataset are available for reproducibility at https://github.com/DarthReca/hydro-chronos.

2 Related Work

Forecasting surface water dynamics requires advancements in satellite monitoring, time-series analysis, spatio-temporal modeling, climate data integration, and interpretability. This section reviews existing literature across these domains, contextualizing the contributions of HydroChronos and our proposed methodology.

2.1 Surface Water Monitoring

Satellite remote sensing revolutionized monitoring surface water extent and dynamics across vast scales and diverse temporal resolutions. Landsat [45] and the Sentinel [13] missions have provided decades of optical imagery, forming the backbone of many surface water mapping efforts. Common water delineation methods include spectral indices like the Normalized Difference Water Index (NDWI) [24] and the Modified NDWI (MNDWI) [47], leveraging water's spectral reflectance. Machine learning classifiers have also been widely employed for more accurate and robust water body extraction [8, 14]. These efforts have culminated in the development of several large-scale and global surface water datasets [28, 48].

These datasets offer insights into past water dynamics but focus on retrospective analysis, not forecasting, providing only the masks of water extents over time. Moreover, they are not explicitly structured for the development and validation of long-term predictive models that integrate auxiliary drivers like climate. Additionally, they are still obtained from automatic extraction, and so their accuracy is dependent on the employed model. HydroChronos is specifically designed for multi-year surface water dynamics forecasting, integrating in a single dataset, imagery, climate variables, and DEM. This multi-modal structure, curated for diverse hydrological systems, provides a rich foundation for developing generalizable models.

2.2 Time-Series Forecasting in Hydrology

The analysis and forecasting of hydrological variables like streamflow or water levels was studied for a long time [23]. Both traditional machine learning approaches [38] and deep learning-based (e.g, based on Long Short-Term Memory [17]) have been applied in Earth observation data [30, 33]. While recent advancements in remote sensing posed the problem of forecasting a snapshot image of the future, including exogenous variables [4], no application can be seen in hydrology, to the best of our knowledge. Forecasting surface water dynamics is influenced by complex, non-linear interactions between past states, seasonality, and external drivers. HydroChronos proposes to fill the gap, enabling scientists to

experiment with a large-scale corpus tailored for hydrological applications based not only on image data, but also on exogenous climate variables.

Deep learning architectures have demonstrated remarkable success in a wide array of environmental modeling and Earth observation tasks, thanks to their ability to learn hierarchical features from large, complex datasets. U-Net architectures [32] still prove to be a strong baseline for semantic segmentation of satellite imagery [7, 10], including water body delineation [8], and more recently, for spatio-temporal forecasting tasks where the output is an image or a sequence of images [42]. Our AquaClimaTempo UNet (ACTU), building on similar architectures[4, 19, 37], adds a dedicated branch for time-series climate data integration and gated fusion to balance climate and optical features. This allows the model to learn how climatic factors modulate surface water dynamics, moving beyond purely auto-regressive image forecasting.

2.3 Explainable AI in Earth Sciences

The complexity of AI models in critical domains like Earth sciences demands transparency and interpretability [2, 43]. XAI techniques have been applied to environmental models to identify key input features or understand model behavior [49]. In our case, the integration of climate variables such as precipitation, temperature, and evapotranspiration is well-established as essential for accurate hydrological modeling [9, 15]. Combining climate data with satellite observations presents significant opportunities but also challenges, including issues of scale mismatch, data assimilation, and capturing complex, potentially lagged, interactions. Prior studies focus primarily on predictive accuracy, often overlooking interpretability. In contrast, our approach incorporates XAI to analyze the relative importance of historical spatio-temporal patterns and climate drivers in predicting future surface water changes.

3 HydroChronos Dataset

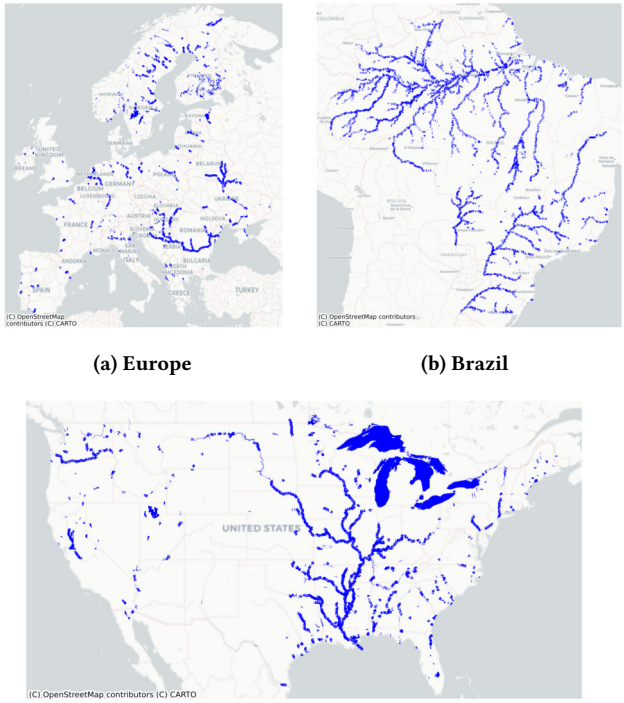
In this section, we present the newly created dataset: HydroChronos. The dataset is composed of time series of images from Landsat-5 and Sentinel-2, time series of climate variables from TERRACLIMATE [1], and a DEM. The selected lakes' and rivers' names are derived from HydroLAKES [25] and HydroRIVERS [20] and cover USA, Europe, and Brazil as shown in Figure 2.

3.1 Landsat-5 and Sentinel-2 images

To capture long-term changes and recent dynamics, HydroChronos uses imagery from Landsat-5 and Sentinel-2. Sentinel-2 provides imagery with superior spatial resolution (10m/20m) and spectral quality compared to Landsat-5 (30m). However, its temporal coverage is limited to the period from 2015 to 2024. To extend the historical perspective, we include Landsat-5 imagery, which covers the period from 1990 to 2010.

To ensure data quality, we selected Top-Of-Atmosphere (TOA) images with the lowest cloud coverage possible, prioritizing clear observations of water bodies. To ensure comparable hydrological conditions and minimize unrelated seasonal variability, we selected May-August images (Northern Hemisphere summer).

Recognizing the spectral differences between the two sensors, we selected a consistent set of spectral bands. Sentinel-2 provides



(a) Europe **(b) Brazil**

(c) United States

Figure 2: Distribution of lakes and rivers in HydroChronos

up to 13 spectral bands, while Landsat-5 offers 7. We harmonized the data by selecting 6 comparable bands that are available on both sensors, as shown in Table 1. All imagery is provided at a spatial resolution of 30m and projected to WGS84. An RGB version of a sample can be seen in Figure 3a.

Table 1: Landsat (L) and Sentinel (S) coupled bands included in the dataset. *NIR* is Near InfraRed and SWIR is *Short-Wave InfraRed*

Landsat	Sentinel	Description	Central Wavelength (L/S)
B1	B2	Blue	485/492 nm
B2	B3	Green	560/560 nm
B3	B4	Red	660/665 nm
B4	B8	NIR	830/833 nm
B5	B11	SWIR	1650/1610 nm
B7	B12	SWIR	2220/2190 nm

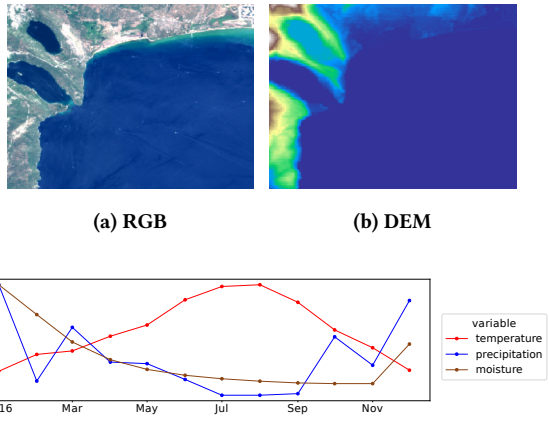
3.2 Digital Elevation Model

A static Digital Elevation Model (DEM) provides essential topographic context for hydrological analysis. The DEM for HydroChronos is sourced from the Copernicus GLO30 DEM [3] dataset, which provides global coverage at a spatial resolution of approximately 30 meters. A sample can be seen in Figure 3b. This single-timestep layer captures the terrain elevation for each study area, crucial

for tasks such as watershed delineation, flow accumulation analysis, and understanding the topographical influence on water body characteristics [26].

3.3 Climate Variables

To complement the remote sensing data with key environmental drivers, HydroChronos includes time series of climate variables from the TERRACLIMATE [1] dataset. It provides monthly climate data globally at a resolution of approximately 4.6 km. The dataset includes 14 variables: actual evapotranspiration, climate water deficit, reference evapotranspiration, precipitation accumulation, runoff, soil moisture, downward surface shortwave radiation, snow water equivalent, maximum temperature, minimum temperature, vapor pressure, vapor pressure deficit, Palmer Drought Severity Index, and wind speed at 10m. We include the complete monthly time series for the periods corresponding to the imagery: 1990-2010 and 2015-2024. A subset of these time series can be seen in Figure 3c. These climate variables can be used to analyze the relationship between climatic conditions and observed changes in water bodies captured by the satellite imagery.



(a) RGB (b) DEM

(c) Climate timeseries

Figure 3: Sample in the three modalities: optical (RGB channels only for visualization), DEM, and climate.

3.4 Splits

Since each water basin has its peculiar behavior, which could be difficult to summarize in simple hydrological variables even if two of them are near, the most straightforward way is to generalize temporally instead of trying to generalize spatially. We use the Landsat-5 dataset (from 1990 to 2010) to pretrain the model; the old sensor (from the 80s) has many sensor errors and noise; however, the huge amount of collected data over a large temporal span makes this sensor ideal to learn dynamics for many areas around the globe. Sentinel-2, which is more modern, has a higher revisit frequency and higher quality images (from 2015 until now), is used for testing and further fine-tuning in the following way: the Brazilian rivers are used for fine-tuning, to align the features learned from Landsat-5 to Sentinel-2 sensor, while Europe and USA are used for the testing. In this way, we collect around 1900 time series for testing and around 16000 for training, for a total of over 100 thousand single images.

4 Tasks

In this section, we delineate the target used in the proposed tasks and how each task is formulated. A visual example of the tasks is shown in Figure 5.

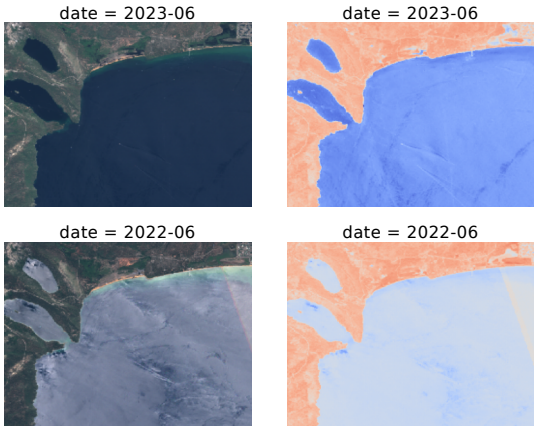


Figure 4: RGB sample at two different timesteps and the corresponding MNDWIs.

4.1 General Target

Given the difficulty in finding yearly annotations of water dynamics all over the world, we based our task on one renowned geo-index to detect water: MNDWI. A similar approach for NDVI was already explored [4]. It is based on the physical properties of water, which is highly reflective in the green channel (G) and absorbs SWIR: $MNDWI = (G - SWIR)/(G + SWIR)$. This approach, while prone to noise, has the advantage of predicting directly the changes regarding physical properties (e.g., icing [41], turbidity [46], pollution [21, 22, 50]) rather than only focusing on water extents and avoiding a costly annotation process. An example can be seen in Figure 4, where the icing process lowers the MNDWI values of the same area.

To avoid any inconsistency, due to the cloud, we apply cloud masking to invalid areas. Additionally, the imperfection of cloud masking and possible sensor errors is avoided with the following setting: given a past timeseries P and a future timeseries F of MNDWIs, our target T is defined as: $T = median(P) - median(F)$, where the median is applied pixelwise over the time axis. In this way, instead of predicting the immediate, possibly noisy future, we target the future trend of the area. This pre-processing step is crucial for a large-scale analysis across diverse regions like the US, Europe, and Brazil, as it effectively reduces localized noise, accounts for minor short-term fluctuations in water levels or atmospheric interference, and smooths the MNDWI signal. Consequently, the subsequent change detection focuses on more persistent and significant alterations in water dynamics rather than ephemeral changes or sensor

artifacts. Since MNDWI $\in \{-1, 1\}$, $T \in \{-2, 2\}$. The target distribution is strongly skewed (the median is 0.01, and the 75th percentile is 0.06) towards zero, as expected.

4.2 Change Detection

The first proposed task is binary change detection, which can be framed as binary semantic segmentation. Given a timeseries P, a target T, and a threshold t to define what we consider a relevant change, we create a binary mask $M_c = |T| > t$. The task focuses on creating a model to predict M_c.

4.3 Direction Classification

This task can be framed as a multiclass semantic segmentation task with 3 classes: negative, positive, or no change. Given a timeseries P, a target T, and a threshold t, we create a mask M_d where a pixel m_d is assigned to the negative change class if $m_d < t$, to the positive change class if $m_d > t$, otherwise it is assigned to the no change class. The task focuses on creating a model to predict M_d.

4.4 Magnitude Regression

The previous tasks assume the existence of a threshold t to define relevant changes. However, it can be of interest to model every "small" change in the area. Given a timeseries P and a target T, the task focuses on creating a model to regress the values of $|T|$. This task can be framed as pixel-wise regression. Preliminary experiments also tried to address the regression of T, but with little success, so we reported only these settings as baseline, leaving this last task for future work.



(a) Regression (b) Change (c) Direction

Figure 5: Visual example of tasks for Lake Tahoe. In regression, the values range from 0 to 2 (blue to red). In change detection, labels are no-change (blue) and change (red). In direction classification, labels are negative change (blue), no-change (grey), and positive change (red).

5 Methodology

In this section, we first discuss our proposed baseline, and then we explain the regression loss we employed.

5.1 AquaClimaTempo UNet

The AquaClimaTempo UNet (ACTU) architecture is depicted in Figure 6. If DEM of shape $1 \times 1 \times W \times H$ is given, it is repeated T times, one for each sample of the image timeseries P to which is concatenated along the channel axis. The image time series of shape $T \times C \times W \times H$ (eventually $C+1$, in case DEM is concatenated) is given to the *Pyramid Image Feature Extractor* (e.g, ConvNext [44]). It processes each image independently, and since it is pyramidal, it provides L features, one for each level, of different shapes from F_0 to F_L. If the climate variables are provided, they are one for each of the T images and cover the past T_1 months of each image. The variables are C_1. The *climate encoder* produces L features with shapes from F_0 to F_L. The *gated fusion* takes as input both the image and climate features at different levels and dynamically balances the contribution, outputting the same L features of shapes from F_0 to F_L, one for each of the T images. The *ConvLSTM* layers flatten the time dimension, producing a representation for the timeseries for each of the L levels of shape from F_0 to F_L. Finally, the *UNet decoder* takes the multilevel features and creates the final prediction mask by concatenating the features created by the expanding path with the ones obtained with the ConvLSTMs.

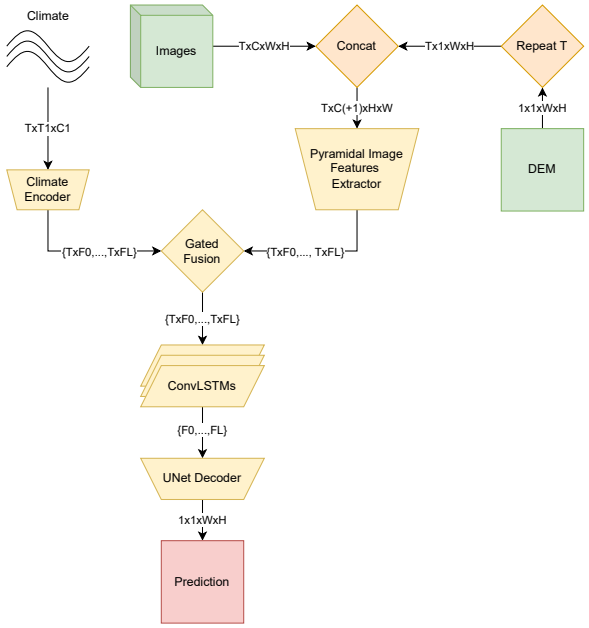


Figure 6: AquaClimaTempo UNet (ACTU) architecture. If DEM is provided, it is repeated once per sample in the image timeseries and concatenated along the channel axis. The *Pyramidal Image Feature Extractor* provides multiscale embeddings. If a climate timeseries is provided, the *climate encoder* provides multiscale embeddings which are *gate fused* with the image embeddings. *ConvLSTMs* provide multiscale embeddings for the timeseries, which are used in the *UNet decoder* to provide the final prediction.

5.1.1 Climate Encoder. The climate encoder processes X_{clim}, one timeseries for each of the T images, with length T_1 and C_1 features. In this way, we can incorporate historical trends with a finer timestep (i.e, monthly instead of yearly). They are independently processed and projected with a linear layer to create F_{proj} (Equation (1)). They are then processed to create K_l representation to spatially match the image features using L blocks composed of an initial Conv2D (with kernel 1) and GELU (Equation (2)), followed

by S series of nearest neighbor upsampling (with resize factor 2), Conv2D (with kernel 3), and GELUs (Equation (3)).

$$F_{proj} = \text{Linear}(\Phi_{\text{LSTM}}(X_{clim})) \tag{1}$$

$$K_l^0 = \text{GELU}(\text{Conv2D}_{1\times1}(F_{proj})) \tag{2}$$

$$K_l^s = \text{GELU}(\text{Conv2D}_{3\times3}(\text{Upsample}_{\times2}(K_l^{s-1}))) \tag{3}$$

5.1.2 Gated Fusion. The climate information can act differently based on the image itself, and also based on the area of the single image. To dynamically adapt the contribution of climate features over the image features, we apply a gate fusion. This solution provides a value in the range 0-1 for each element of the matrix to weight the contribution of optical and climate features dynamically. The weights are obtained by concatenating along the channels axis climate and optical features, and applying a Conv2D (with kernel 3), a ReLU, and Conv2D (with kernel 1), and finally a sigmoid to constrain the input in the range 0-1. Given the climate feature K_l and the corresponding image feature I_l at the level L, the gated fusion can be formulated:

$$Z = \text{Concat}(K_l, I_l) \tag{4}$$

$$\alpha = \sigma(\text{Conv2D}_{1\times1}(\text{ReLU}(\text{Conv2D}_{3\times3}(Z)))) \tag{5}$$

$$F_l = \alpha I_l + (1-\alpha)K_l \tag{6}$$

Each F_l is then used in the decoder the make the final prediction.

5.2 Regression Loss

Since the satellite resolution can vary and the problem shows a strong imbalance, L1 or L2 losses alone can be insufficient, as they are also sensitive to noise. Our regression loss makes use of Huber-Loss as a starting point, but combines a multi-scale approach with a wavelet decomposition.

5.2.1 Multiscale Loss. The multiscale loss L_{MS}, given a regression loss L, prediction P, and ground truth T can be defined as:

$$L_{MS}(P,T) = \frac{1}{M}\left(L(P,T) + \sum_{i=1}^{M} L(D_{s_i}(P), D_{s_i}(T))\right) \tag{7}$$

where $S = \{s_0, ..., s_M\}$ are M different scale factors and D_x is the downscale operation with factor x.

5.2.2 Wavelet Loss. The Discrete Wavelet Transform (DWT) decomposes both the prediction and the target into different frequency sub-bands, instead of directly comparing pixel values in the spatial domain. Compared to the Fourier transform, it is computationally more efficient ($O(N)$ compared to $O(N\log(N))$). This allows the loss to penalize errors at different scales and orientations (horizontal, vertical, diagonal). DWT decomposes the image with two coefficients: the approximation coefficients Y_L (the low-frequency components, capturing its coarse structure) and detail coefficients Y_H (the high-frequency components at N levels and horizontal, vertical, diagonal orientations, capturing finer details). Given a regression loss L, prediction wavelet coefficients Y_H^p and Y_L^p, and ground truth coefficients Y_H^t and Y_L^t, the wavelet loss L_W can be defined as:

$$L_L(Y_L^p, Y_L^t) = \text{mean}(L(Y_L^p, Y_L^t)) \tag{8}$$

$$L_{H,i}(Y_{H,i}^p, Y_{H,i}^t) = \text{mean}(L(Y_{H,i}^p, Y_{H,i}^t)) \tag{9}$$

$$L_W(Y_L^p, Y_L^t, Y_H^p, Y_H^t) = \alpha L_L(Y_L^p, Y_L^t) + \sum_{i=1}^{N} w_i \cdot L_{H,i}(Y_{H,i}^p, Y_{H,i}^t) \tag{10}$$

where α defines the weight of the low-frequency loss and $W = \{w_0, ..., w_N\}$ the weights of the high frequency losses.

The final loss is a weighted mean of the multiscale and wavelet losses: $L_T = \alpha L_{MS} + (1-\alpha)L_W$.

5.3 Explainable AI analysis

The Explainable AI analysis investigates model behavior from two complementary perspectives. First, we leverage climate-derived attributes to perform *subgroup discovery and feature attribution*, aiming to reveal systematic performance disparities across interpretable climate conditions. Second, we conduct a *per-channel saliency analysis* to quantify the contribution of individual input modalities to the model's predictions.

5.3.1 Climate Subgroup Discovery and Feature Attribution. Model performance in spatial machine learning can vary significantly across different environmental conditions. Aggregated metrics may conceal systematic failures concentrated in specific climatic regimes. To uncover and explain such disparities, we adopt a post hoc analysis framework based on subgroup discovery and feature attribution.

Climate Subgroup Discovery. Let $\mathcal{D} = \{(x_i, y_i, \hat{y}_i)\}_{i=1}^N$ denote the evaluation dataset, where $x_i \in \mathcal{X}$ represents the input sample, $y_i \in \mathcal{Y}$ is the ground-truth label, and $\hat{y}_i \in \mathcal{Y}$ is the model prediction. Each x_i is associated with a set of interpretable climate-derived attributes $A(x_i) = \{a_1, \ldots, a_k\}$ obtained via feature binning.

We define a *subgroup* $S \subseteq \mathcal{D}$ as the set of samples satisfying a conjunction of attribute-value conditions:

$$S = \{x_i \in \mathcal{D} \mid a_j(x_i) = v_j \quad \forall j \in J\},$$

where $J \subseteq \{1, \ldots, k\}$ indexes selected attributes and v_j denotes specific bin values. To evaluate the behavior of a model over S, we use the notion of subgroup performance divergence [27] defined as:

$$\Delta_m(S) = m(S) - m(\mathcal{D}),$$

where $m : \mathcal{D} \to \mathbb{R}$ is a scalar performance metric (e.g., precision, recall), $m(S)$ is the metric evaluated over the subgroup, and $m(\mathcal{D})$ is the global reference over the entire dataset.

We automatically identify the subgroups with large divergence scores using `DivExplorer` [27], which enumerates statistically significant subgroups under a minimum support constraint θ. This process identifies climate conditions under which the model under- or over- performs.

Feature Attribution. To explain which features contribute most to these performance deviations, we use the notion of Global Shapley values [27]. The Global Shapley value is a generalization of the Shapley value [35] which estimates the contribution of each attribute-value to the divergence across all identified subgroups above the support constraint. The higher the value, the more the

attribute-value term contributes to the divergence in performance. A positive contribution of a term indicates that it is associated with a performance metric m higher than the average on the overall dataset. We refer the reader to [27] for its formal definition.

5.3.2 Per-channel Saliency. A central goal in XAI is to identify which input components most significantly influence a model's predictions. Beyond interpretability, this analysis has practical implications, such as reducing computational overhead by pruning less informative input channels. A common approach for estimating input relevance is perturbation-based and involves perturbing the input and measuring the resulting change in the model's output. Perturbation-based techniques [5, 11, 51] provide a straightforward method for attributing importance scores to input dimensions based on their effect on model behavior. We adopt a perturbation-based strategy to compute *per-channel saliency*, aimed at quantifying the relevance of each input channel. We adapted the method proposed in [29] to our time-series setting. Let $x \in \mathbb{R}^{T\times C\times H\times W}$ denote the input tensor, where T is the number of temporal frames, C the number of channels, and $H \times W$ the spatial resolution. For a given test sample, we first compute the model's baseline prediction $\hat{y} = \mathcal{M}(x, \text{DEM}, \text{Climate})$, and evaluate it using a suite of performance metrics. To assess the importance of channel $c \in \{1, \ldots, C\}$, we generate a perturbed input $x^{(-c)}$ by zeroing out the c-th channel across all time steps. We then recompute the model output as $\hat{y}^{(-c)} = \mathcal{M}(x^{(-c)}, \text{DEM}, \text{Climate})$. The saliency of channel c is defined as the change in a given performance metric m:

$$\Delta m_c = m(\hat{y}, y) - m(\hat{y}^{(-c)}, y),$$

where y is the ground truth. A larger Δm_c indicates that channel c has a greater influence on model performance, since its absence leads to a more substantial degradation in prediction quality. We then average the per-channel saliency scores across the test set to obtain a dataset-level contribution. This process is extended to the DEM input, which is ablated entirely to measure its global contribution. Overall, this saliency analysis provides both local (per-sample) and global (dataset-level) interpretability, helping to identify which input modalities most influence the model decisions in spatiotemporal tasks.

6 Experimental Results

In this section, we present the experimental settings and the results compared to simple statistical methods, namely constant prediction and persistence. Constant prediction is simply predicting that no change will happen in the future. Persistence for robustness is computed as the difference between the last known timestep and the median of the previous (thresholded with t for classification).

6.1 Experimental Settings

ACTU is pretrained for 50 epochs on the Landsat subset, and fine-tuned for 20 epochs on Sentinel-2 with a cosine decaying learning rate scheduler with a 5% warmup. The maximum learning rate is 5e-4 for the pretraining and 5e-6 for the fine-tuning. The batch size was set to 8. The loss for classification is a combo loss composed of generalized dice and focal losses. The vision backbone for the encoder is ConvNextV2 [44], in base (ACTU) and large (ACTU-L) versions. The length of the two image time series was set to 5. To avoid overwhelming the neural network with information (and deal with eventual multicollinearity and increasing costs), we select a subset of 5 climate variables (maximum temperature (tmmx), actual evapotranspiration (aet), runoff (ro), precipitation (pr), and soil moisture (soil)) that should be relevant for the task [6, 16, 31]. We also use these five variables for the analysis in Section 5.3.1. Since this analysis requires categorical attributes to define subgroups, we discretized each variable into three interpretable bins. Specifically, we computed the standard deviation of each climate variable over the input time series to quantify temporal variability. These variability scores were discretized by frequency into three levels—*low (L)*, *medium (M)*, and *high (H)*—thus enabling the construction of climate subgroups with distinct fluctuation profiles. In our experiments, we threshold at $t = 0.1$ (i.e., ~85th percentile) to remove possible sensor noises, atmospheric effects, phenological changes, and coregistration errors. A key requirement for training a neural network is the generation of a consistent change mask across the entire diverse study area. A fixed threshold ensures that the definition of 'change' is uniform. While this approach is necessary for large-scale change detection, it is acknowledged that the precise magnitude of MNDWI difference that corresponds to a 'relevant' real-world change can still exhibit some variability due to the diverse nature of water bodies, atmospheric conditions, and local landscape characteristics across the US, Europe, and Brazil.

6.2 Evaluation Metrics

We evaluate classification tasks using precision (P), recall (R), and F1-score (F) for each class. Since the regression problem is pixel-wise, and many areas have values near zero, it can be considered "unbalanced". We evaluate the regression quality with Mean Absolute Error (MAE) and the Pearson Correlation (PC). We also compute MAE on the top-10 (MAE@10) and top-20 (MAE@20) highest valued pixels to account for the imbalance of the values. We threshold the regressed values at $t = 0.1$ (as for classification) and $t = 0.2$, where we compute precision (P@t), recall (R@t), and F1-score (F@t). This allows us to assess the trade-offs between detecting relevant pixels and the accuracy of those detections at different sensitivity levels.

6.3 Change Detection

Table 2 presents the binary change detection results, highlighting the performance of various model configurations. All ACTU model variants demonstrate a substantial and statistically significant improvement in detecting changes (CHG) compared to the *Constant* and *Persistence* baselines. For instance, the baselines achieve F1-scores of 0 and 34.98, respectively, whereas all ACTU configurations surpass an F1-score of 45 for this class. The inclusion of climate variables (C) with the base ACTU model improves the no-change class (NoCHG) F1-score but results in a slight decrease in the CHG F1-score. Conversely, incorporating only DEM data (D) enhances the CHG F1-score, the highest among the standard ACTU variants for this metric, while also slightly improving NoCHG performance. When both DEM and climate data are utilized, the model achieves the highest CHG Recall (62.33), proving its superior capability in identifying actual change instances, though its F1-score (48.67) is marginally lower than the DEM-only configuration. The larger

backbone model, ACTU-L, shows a modest improvement in CHG precision. Generally, a larger backbone does not seem to provide consistent improvements over the base version.

Table 2: Change detection results for models optionally using DEM (D) and climate variables (C). * indicates statistically significant difference (p < 0.01) with respect to persistence according to the t-test. ° indicates the statistical difference comparing ACTU-L with the same configuration of ACTU.

			No Change (NoCHG)			Change (CHG)		
Model	D	C	P	R	F	P	R	F
Constant	N	N	81.54	**100**	**89.25**	0	0	0
Persistence	N	N	88.73	41.64	54.07	23.86	**81.77**	34.98
ACTU	N	N	90.51*	82.78*	85.66*	44.87*	60.65*	48.79*
ACTU	N	Y	88.92*	85.86*	86.6*	45.45*	53.1*	45.83*
ACTU	Y	N	**90.57**	82.89*	85.75*	45.19*	61.01*	**49.38***
ACTU	Y	Y	90.53*	81.71*	85.08*	43.68*	62.33*	48.67*
ACTU-L	N	N	90.03°	84.3°	86.43°	**45.46°**	57.34°	47.94°
ACTU-L	Y	Y	89.2°	84.95°	86.37°	45.03°	54.39°	46.49°

6.4 Direction Classification

Table 3 details the performance for the direction classification task, categorizing changes into negative (NEG), no change (NONE), or positive (POS). All ACTU model variants achieve statistically significant and considerable improvements over the *Constant* and *Persistence* baselines. This is particularly evident for POS and NEG, where the *Persistence* achieves 17.48 and 11.61, respectively, compared to ACTU 30.27 and 19.47. All models excel at identifying NONE, with ACTU variants consistently reaching F1-scores around 86-87. However, accurately classifying the direction of change (POS and NEG) is inherently more challenging, as reflected by their lower F1-scores compared to NONE across all models. Introducing climate variables notably improves the F1-score for POS, though it slightly reduces performance for NEG. Conversely, adding DEM data alone does not yield a clear F1-score improvement for either change direction class, slightly decreasing NEG and POS F1-scores. The combined use of DEM and climate data results in a robust NONE F1-score (87.6) and a POS F1-score (20.77). The larger ACTU-L model shows modest F1 improvements for NONE (87.76) and POS (20.88) over its smaller counterpart, but a decrease for NEG. These results suggest that while ACTU is strongest for detecting negative changes, incorporating climate data is particularly beneficial for identifying positive changes. The overall task of precise direction classification remains complex, with input data types showing varied impacts on different change categories.

6.5 Magnitude Regression

Table 4 details the magnitude regression performance, where all ACTU model variants demonstrate statistically significant and substantial improvements over the *Constant* and *Persistence* baselines across all reported metrics. The standard ACTU model achieves a low MAE of 0.0261 and a PC of 46.45. While the incorporation of DEM (D), climate (C) variables, or both tends to slightly increase overall MAE and decrease PC, the inclusion of DEM markedly improves the F1-scores when regression outputs are thresholded to identify significant changes. Specifically, F@0.1 and F@0.2 improve due to enhanced recall. This indicates that while the base model excels at general magnitude prediction, DEM input is particularly beneficial for more accurately identifying pixels undergoing substantial change. The larger backbone model, ACTU-L, emerges as the top-performing configuration. It improves MAE for the top 10% and 20% highest magnitude changes and achieves the highest F1-scores for thresholded significant changes. Although its overall MAE is marginally higher than ACTU, its PC is slightly better. In contrast, ACTU-L with climate and dem, while improving general MAE and PC over its smaller counterpart ACTU, does not reach the thresholded performance levels of ACTU-L.

6.6 Ablation Studies on regression loss

To understand the contribution of the proposed composed loss L_T, we report in Table 5 the comparison with its loss components alone and the employment of a standard regression loss (Huber loss, the same used in the other derivative losses). We compare the ACTU without any additional information for simplicity. Comparing the Huber loss (L) to the multiscale version (L_{MS}), L_{MS} provides better recalls and so better F1-scores, and lower values of MAEs in the top-highest values. The wavelet loss (L_W) provides good performance in regression, but struggles with classification metrics. This is probably due to the frequency domain, which lacks any spatial information. The linear combination L_T proves its benefits in regression metrics when looking at MAE and PC (at least +2% improvement). Looking at classification metrics, it enhances the precision while affecting the recall. The F@0.1 remains not significantly affected by the recall loss, while F@0.2 is enhanced. It can be easily seen that it blends the highest recall of spatial loss (L_{MS}), with the highest precision of frequency loss (L_W), providing a balanced contribution of both. This loss proved to be a good alternative; still, a more extensive search could be performed in the future to select even better hyperparameters.

7 Analysis and Discussion

In this section, we present the insights derived from the XAI analysis. We begin with the Climate Subgroup Discovery and Feature Attribution, which shows how the model's performance changes according to climate variations and highlights its reliance on specific climate variables. We then examine the Per-channel Saliency to assess the relative importance of the individual spectral bands and DEM.

7.1 Climate Subgroups and Feature Attribution

Climate Subgroups. We identify climate subgroups that consistently challenge the model across different tasks. These understandings enable us to outline critical samples characterized by hard-to-learn environmental patterns. We first perform the subgroup discovery over the five climate variables, discretized according to their intra-series variability. For each task and class, we extract the subgroup exhibiting the highest divergence, as detailed in Appendix B. We then conduct a comparative analysis of these subgroups by

Table 3: Direction classification results for models optionally using DEM (D) and climate variables (C). * indicates statistically significant difference (p < 0.05) with respect to persistence according to the t-test. ° indicates the statistical difference comparing ACTU-L with the same configuration of ACTU.

			Negative Change (NEG)			No Change (NONE)			Positive Change (POS)		
Model	D	C	P	R	F	P	R	F	P	R	F
Constant	N	N	0	0	0	81.54	**100**	**89.25**	0	0	0
Persistence	N	N	14.94	47.37	17.48	88.73	41.64	54.07	9.25	**31.18**	11.61
ACTU	N	N	27.5*	**49.38***	**30.27***	**90.23***	84.58*	86.67*	27.73*	19.84*	19.47*
ACTU	N	Y	**29.84***	36.53*	27.75*	88.44	87.5*	87.42*	26.16*	24.12*	**21.38***
ACTU	Y	N	28.18*	34.43*	25.81*	87.73*	88.65*	87.54*	27.39*	20.25*	19.09*
ACTU	Y	Y	28.36*	37.35*	27.28*	88.18*	88.17*	87.6*	27.98*	22.06*	20.77*
ACTU-L	N	N	28.42°	38.21°	27.62°	88.5°	88.01°	87.76°	**28.25°**	22.04°	20.88°
ACTU-L	Y	Y	27.52°	34.84°	25.31°	88.15	88.09	87.55	27.13°	21.28°	19.44°

Table 4: Magnitude regression results for models optionally using DEM (D) and climate variables (C). * indicates statistically significant difference (p < 0.05) with respect to persistence according to the t-test. ° indicates the statistical difference comparing ACTU-L with the same configuration of ACTU.

Model	D	C	MAE	MAE@10	MAE@20	PC
Constant	N	N	.0351	.142	.1038	-
Persistence	N	N	.1281	.1892	.171	31.81
ACTU	N	N	**.0261***	.0873*	.0611*	46.45*
ACTU	N	Y	.0266*	.0911*	.0639*	44.4*
ACTU	Y	N	.0297*	.0886*	.0643*	44.46*
ACTU	Y	Y	.0315*	.088*	.0634*	41.41*
ACTU-L	N	N	.0275°	**.0843°**	**.0589°**	**46.62**
ACTU-L	Y	Y	.0282°	.0923°	.066°	43.45°

Model	D	C	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2	F@0.2
Constant	N	N	0	0	0	0	0	0
Persistence	N	N	23.86	**81.77**	34.98	13.38	**75.72**	20.25
ACTU	N	N	**51.97***	41.61*	43.04*	45.27*	24.48*	26.85*
ACTU	N	Y	51.32*	41.59*	42.77*	42.34*	18.12*	21.02*
ACTU	Y	N	49.36*	45.09*	43.64*	40.67*	28.26*	27.8*
ACTU	Y	Y	46.92*	45.18*	42.67*	39.57*	27.91*	27.3*
ACTU-L	N	N	50°	**47.19°**	**45.61°**	44.45°	27.55°	**28.65°**
ACTU-L	Y	Y	51.63°	38.48°	40.65°	43.24°	21.41°	23.41°

inspecting the samples they encompass, in order to uncover recurring problematic regions across tasks. We show the findings in Table 6. Notably, there is strong agreement across all the tasks in the difficulty of predicting that the area will change in the regions of Great Salt Lake and Utah Lake, as they appear in the worst-performing subgroups for both change, direction, and regression models. Conversely, Rainy Lake and Woods Lake are particularly problematic when the model attempts to predict stable conditions (i.e., no change), suggesting that temporal climate fluctuations in

Table 5: Comparison between the wavelet loss (L_W), multi-scale loss (L_{MS}), their combination L_T, and standard application of the regression loss (L). * indicates statistically significant difference (p < 0.01) with respect to L_T according to the t-test.

Model	MAE	MAE@10	MAE@20	PC
L_T	**.0261**	.0873	.0611	**46.45**
L	.029*	.0861*	.06*	42.9*
L_{MS}	.0291*	**.0839***	**.0585***	43.82*
L_W	.0285*	.1047*	.0765*	42.1*

Model	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2	F@0.2
L_T	**51.97**	41.61	43.04	**45.27**	24.48	**26.85**
L	46.14*	45.65*	42.25	39.89*	24.25	24.61*
L_{MS}	45.85*	**49.84***	**44.83**	39.38*	**27.2**	26.75*
L_W	51.24	25.11*	29.88*	38.51*	16.24*	18.42*

Table 6: Lakes associated with the worst-performing climate subgroups across the three tasks: change detection (C), direction classification (D), and regression (R), with the affected classes (for C and D) and MAE metric for R.

Lake	(C)	(D)	(R)
Great Salt Lake	CHG	POS	MAE
Utah Lake	CHG	POS	MAE
Rainy Lake	NoCHG	NONE	—
Woods Lake	NoCHG	NONE	—
Lake Texoma	—	POS	MAE
Pyramid Lake	—	POS	MAE

these areas might mimic weak change signals. Additionally, Lake Texoma and Pyramid Lake consistently appear in subgroups associated with poor performance in both the regression task and the

detection of positive changes, pointing to a possible shared climate signal that the model struggles to generalize across these scenarios.

Feature Attribution. We compute the Global Shapley values to quantify the contribution of each attribute-value pair to the overall divergence. This allows us to identify which factors are most responsible for performance variations across subgroups. Table 7 summarizes the two *Most* and *Least* contributing levels of climate variability for each of the three tasks. In the change detection task, precipitation (pr) and soil moisture (soil), particularly under conditions of low or high variability, consistently emerge as strong contributors to model performance. Maximum temperature (tmmx) also appears repeatedly across all tasks, with its variability positively correlated with higher prediction accuracy. In contrast, high variability in evapotranspiration (aet) is frequently among the least contributing factors, indicating that it may introduce instability or ambiguity that the model struggles to effectively capture. These findings indicate that model performance is not uniformly distributed across climatic regimes and that temporal variability in certain features (e.g., soil, aet) can lead to systematic failure modes. Identifying failure-prone climate subgroups not only enhances our understanding of climate feature relevance but also provides actionable insights for improving model robustness through targeted data augmentation, domain adaptation, and climate-aware validation strategies.

7.2 Per-channel Saliency

We perform a per-channel saliency analysis to evaluate the contribution of each input modality. For change detection (C) and direction classification (D), we computed the average drop in F1 score resulting from the ablation of individual channels. For the regression task (R), we instead measured the change in Mean Absolute Error (MAE) and Pearson Correlation (PC). We summarize the saliency scores in Figure 7. All the saliency scores are row-normalized between -1 (confusing channel) and 1 (important channel) to allow for consistent visual comparison, where 0 indicates an irrelevant channel. For the MAE, we normalized its negative value to have a consistent interpretation. The models evaluated included the DEM as an additional input.

In the change detection task, NIR and SWIR channels stand out as the most informative for detecting change events (C-CHG, second row), while RGB channels play a greater role in predicting areas with no change (C-NoCHG, first row). This pattern largely holds in the direction classification task as well. Interestingly, although the NIR channel supports detection of negative changes (D-NEG), it appears to hinder the identification of positive changes (D-POS), revealing a class-dependent interaction with this spectral band. In the regression task, performance improves when the model has access to the full spectrum of channels. For R-PC, NIR and SWIR maintain their prominence, yet the RGB channels also contribute consistently, suggesting a beneficial multispectral synergy.

This analysis highlights the distinct and complementary roles of spectral bands across tasks. NIR and SWIR are crucial for detecting dynamic changes and maintaining high correlation with ground truth signals, while RGB channels remain essential for stable predictions and class discrimination. Moreover, the varying impact of NIR on different direction classes underscores the importance of task-specific channel sensitivity when designing interpretable Earth observation models.

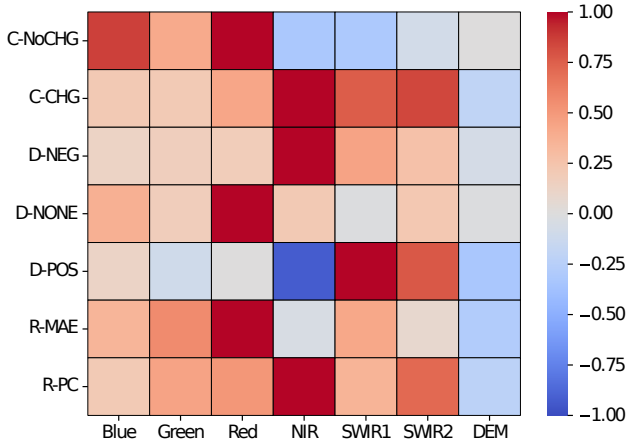


Figure 7: Row-Normalized Per-Channel Saliency related to change detection (C) and direction classification (D), F1-Score for each class, and to MAE and Pearson Correlation (PC) of regression (R).

Table 7: Feature Attribution through Global Shapley values. For each of the tasks, we report the two best and worst contributing climate values (H=high, M=medium, L=low).

		Most Contributing		Least Contributing	
C	NoCHG	pr=H	soil=H	ro=H	aet=H
	CHG	ro=M	soil=L	soil=H	aet=M
D	NEG	pr=L	tmmx=L	pr=M	pr=H
	NONE	pr=H	soil=H	ro=H	aet=H
	POS	tmmx=L	aet=L	ro=H	aet=H
R	MAE	soil=H	tmmx=H	ro=H	aet=H
	PC	tmmx=L	tmmx=H	aet=M	soil=H

8 Conclusion

Our findings lay a foundation for advancing predictive spatio-temporal modeling in hydrology to foster research in water resource management in an era of significant environmental change. Our dataset provides the first large-scale effort to map water evolution over the years, and ACTU provides a baseline that integrates visual features and climate variables. XAI analysis identified salient spectral bands and the impacts of climate variables, providing insight into the flaws and strengths of our model, guiding future work. In future works, additional tasks (e.g., MNDWI regression), models for extreme events, and detailed analysis of the relations between climate and visual features should be investigated. Domain adaptation should be analyzed in detail to apply models to new geographic areas. Integrating data from radar altimetry, such as SWOT, could provide water surface elevation, and in situ measurements of river gauges can complement the spectral data.

References

[1] John T. Abatzoglou, Solomon Z. Dobrowski, Sean A. Parks, and Katherine C. Hegewisch. 2018. TerraClimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958–2015. *Scientific Data* 5, 1 (Jan. 2018). doi:10.1038/sdata.2017.191

[2] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. doi:10.1109/ACCESS.2018.2870052

[3] European Space Agency. 2022. Copernicus DEM. doi:10.5270/esa-c5d3d65

[4] Vitus Benson, Claire Robin, Christian Requena-Mesa, Lazaro Alonso, Nuno Carvalhais, José Cortés, Zhihan Gao, Nora Linscheid, Mélanie Weynants, and Markus Reichstein. 2024. Multi-modal Learning for Geospatial Vegetation Forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 27788–27799.

[5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.

[6] Luca Brocca, Luca Ciabatta, Christian Massari, Stefania Camici, and Angelica Tarpanelli. 2017. Soil Moisture for Hydrological Applications: Open Questions and New Opportunities. *Water* 9, 2 (Feb. 2017), 140. doi:10.3390/w9020140

[7] Daniele Rege Cambrin, Luca Colomba, and Paolo Garza. 2023. CaBuAr: California burned areas dataset for delineation [Software and Data Sets]. *IEEE Geoscience and Remote Sensing Magazine* 11, 3 (Sept. 2023), 106–113. doi:10.1109/mgrs.2023.3292467

[8] Huidong Cao, Yanbing Tian, Yanli Liu, and Ruihua Wang. 2024. Water body extraction from high spatial resolution remote sensing images based on enhanced U-Net and multi-scale information fusion. *Scientific Reports* 14, 1 (July 2024). doi:10.1038/s41598-024-67113-7

[9] Francis H. S. Chiew and Thomas A. McMahon. 2002. Modelling the impacts of climate change on Australian streamflow. *Hydrological Processes* 16, 6 (March 2002), 1235–1245. doi:10.1002/hyp.1059

[10] Isaac Corley, Caleb Robinson, and Anthony Ortiz. 2024. A Change Detection Reality Check. *arXiv preprint arXiv:2402.06994* (2024).

[11] Ian C. Covert, Scott Lundberg, and Su-In Lee. 2021. Explaining by removing: a unified framework for model explanation. *J. Mach. Learn. Res.* 22, 1, Article 209 (Jan. 2021), 90 pages.

[12] Abhirup Dikshit and Biswajeet Pradhan. 2021. Explainable AI in drought forecasting. *Machine Learning with Applications* 6 (Dec. 2021), 100192. doi:10.1016/j.mlwa.2021.100192

[13] M. Drusch, U. Del Bello, S. Carlier, O. Colin, V. Fernandez, F. Gascon, B. Hoersch, C. Isola, P. Laberinti, P. Martimort, A. Meygret, F. Spoto, O. Sy, F. Marchese, and P. Bargellini. 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. *Remote Sensing of Environment* 120 (May 2012), 25–36. doi:10.1016/j.rse.2011.11.026

[14] Gudina L. Feyisa, Henrik Meilby, Rasmus Fensholt, and Simon R. Proud. 2014. Automated Water Extraction Index: A new technique for surface water mapping using Landsat imagery. *Remote Sensing of Environment* 140 (Jan. 2014), 23–35. doi:10.1016/j.rse.2013.08.029

[15] Peter H. Gleick. 1987. The development and testing of a water balance model for climate impact assessment: Modeling the Sacramento Basin. *Water Resources Research* 23, 6 (June 1987), 1049–1061. doi:10.1029/wr023i006p01049

[16] Sara Habibi and Saeed Tasouji Hassanpour. 2025. An Explainable Machine Learning Framework for Forecasting Lake Water Equivalent Using Satellite Data: A 20-Year Analysis of the Urmia Lake Basin. *Water* 17, 10 (May 2025), 1431. doi:10.3390/w17101431

[17] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (Nov. 1997), 1735–1780. doi:10.1162/neco.1997.9.8.1735

[18] Anne Jones, Julian Kuehnert, Paolo Fraccaro, Ophélie Meuriot, Tatsuya Ishikawa, Blair Edwards, Nikola Stoyanov, Sekou L. Remy, Kommy Weldemariam, and Solomon Assefa. 2023. AI for climate impacts: applications in flood risk. *npj Climate and Atmospheric Science* 6, 1 (June 2023). doi:10.1038/s41612-023-00388-1

[19] Hamid Kamangir, Brent S. Sams, Nick Dokoozlian, Luis Sanchez, and J. Mason Earles. 2024. Large-scale spatio-temporal yield estimation via deep learning using satellite and management data fusion in vineyards. *Computers and Electronics in Agriculture* 216 (Jan. 2024), 108439. doi:10.1016/j.compag.2023.108439

[20] Bernhard Lehner and Günther Grill. 2013. Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes* 27, 15 (April 2013), 2171–2186. doi:10.1002/hyp.9740

[21] Yingcheng Lu, Jing Shi, Chuanmin Hu, Minwei Zhang, Shaojie Sun, and Yongxue Liu. 2020. Optical interpretation of oil emulsions in the ocean – Part II: Applications to multi-band coarse-resolution imagery. *Remote Sensing of Environment* 242 (June 2020), 111778. doi:10.1016/j.rse.2020.111778

[22] Yingcheng Lu, Jing Shi, Yansha Wen, Chuanmin Hu, Yang Zhou, Shaojie Sun, Minwei Zhang, Zhihua Mao, and Yongxue Liu. 2019. Optical interpretation of oil emulsions in the ocean – Part I: Laboratory measurements and proof-of-concept with AVIRIS observations. *Remote Sensing of Environment* 230 (Sept. 2019), 111183. doi:10.1016/j.rse.2019.05.002

[23] Deepesh Machiwal and Madan Kumar Jha. 2012. *Hydrologic Time Series Analysis: Theory and Practice*. Springer Netherlands. doi:10.1007/978-94-007-1861-6

[24] S. K. McFEETERS. 1996. The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal of Remote Sensing* 17, 7 (May 1996), 1425–1432. doi:10.1080/01431169608948714

[25] Mathis Loïc Messager, Bernhard Lehner, Günther Grill, Irena Nedeva, and Oliver Schmitt. 2016. Estimating the volume and age of water stored in global lakes using a geo-statistical approach. *Nature Communications* 7, 1 (Dec. 2016). doi:10.1038/ncomms13603

[26] Julio Novoa, Karem Chokmani, Rody Nigel, and Philippe Dufour. 2015. Quality assessment from a hydrological perspective of a digital elevation model derived from WorldView-2 remote sensing data. *Hydrological Sciences Journal* 60, 2 (Jan. 2015), 218–233. doi:10.1080/02626667.2013.875179

[27] Eliana Pastor, Andrew Gavgavian, Elena Baralis, and Luca de Alfaro. 2021. How divergent is your data? *Proceedings of the VLDB Endowment* 14, 12 (2021), 2835–2838.

[28] Jean-François Pekel, Andrew Cottam, Noel Gorelick, and Alan S. Belward. 2016. High-resolution mapping of global surface water and its long-term changes. *Nature* 540, 7633 (Dec. 2016), 418–422. doi:10.1038/nature20584

[29] Daniele Rege Cambrin, Eleonora Poeta, Eliana Pastor, Tania Cerquitelli, Elena Baralis, and Paolo Garza. 2025. KAN You See It? KANs and Sentinel for Effective and Explainable Crop Field Segmentation. In *European Conference on Computer Vision*. Springer, 115–131.

[30] Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. 2019. Deep learning and process understanding for data-driven Earth system science. *Nature* 566, 7743 (Feb. 2019), 195–204. doi:10.1038/s41586-019-0912-1

[31] Michael E. Ritter. 2011. *The Physical Environment: an Introduction to Physical Geography*. Earth Online Media. https://www.earthonlinemedia.com/ebooks/tpe_3e/title_page.html

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. Springer International Publishing, 234–241. doi:10.1007/978-3-319-24574-4_28

[33] Marc Rußwurm and Marco Körner. 2017. Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 1496–1504. doi:10.1109/CVPRW.2017.193

[34] Sukanya S and Sabu Joseph. 2023. *Climate change impacts on water resources: An overview*. Elsevier, 55–76. doi:10.1016/b978-0-323-99714-0.00008-x

[35] Lloyd S Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games II*, Harold W. Kuhn and Albert W. Tucker (Eds.). Princeton University Press, Princeton, 307–317.

[36] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai kin Wong, and Wang chun Woo. 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. arXiv:1506.04214 [cs.CV] https://arxiv.org/abs/1506.04214

[37] Shuting Sun, Lin Mu, Lizhe Wang, and Peng Liu. 2022. L-UNet: An LSTM Network for Remote Sensing Image Change Detection. *IEEE Geoscience and Remote Sensing Letters* 19 (2022), 1–5. doi:10.1109/LGRS.2020.3041530

[38] Jan Verbesselt, Rob Hyndman, Glenn Newnham, and Darius Culvenor. 2010. Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment* 114, 1 (Jan. 2010), 106–115. doi:10.1016/j.rse.2009.08.014

[39] Qinqin Wang, Yuwei Liu, Guofeng Zhu, Siyu Lu, Longhu Chen, Yinying Jiao, Wenmin Li, Wentong Li, and Yuhao Wang. 2025. Regional differences in the effects of atmospheric moisture residence time on precipitation isotopes over Eurasia. *Atmospheric Research* 314 (March 2025), 107813. doi:10.1016/j.atmosres.2024.107813

[40] Xander Wang and Lirong Liu. 2023. The Impacts of Climate Change on the Hydrological Cycle and Water Resource Management. *Water* 15, 13 (June 2023), 2342. doi:10.3390/w15132342

[41] Stephen G. Warren. 2019. Optical properties of ice and snow. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 377, 2146 (April 2019), 20180161. doi:10.1098/rsta.2018.0161

[42] Jonathan A. Weyn, Dale R. Durran, and Rich Caruana. 2020. Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems* 12, 9 (Sept. 2020). doi:10.1029/2020ms002109

[43] Samek Wojciech, Montavon Grégoire, Vedaldi Andrea, Kai Hansen Lars, and Müller Klaus-Robert (Eds.). 2019. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing. doi:10.1007/978-3-030-28954-6

[44] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. 2023. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16133–16142. doi:10.1109/CVPR52729.2023.01548

Table 8: Most divergent climate subgroups for each task and class with associated divergence scores indicating the severity of model underperformance.

Task	Class	Most Divergent Subgroup	Divergence Δ
D	NEG	{pr_bin=H, tmmx_bin=H, ro_bin=L}	-0.24
D	NONE	{aet_bin=H, tmmx_bin=H, soil_bin=L, pr_bin=M, ro_bin=H}	-0.18
D	POS	{soil_bin=M, pr_bin=M, aet_bin=L, tmmx_bin=M}	-0.14
C	NoCHG	{soil_bin=L, pr_bin=M, ro_bin=H, aet_bin=H, tmmx_bin=H}	-0.18
C	CHG	{tmmx_bin=M, ro_bin=L, pr_bin=L, soil_bin=M, aet_bin=M}	-0.22
R	MAE	{tmmx_bin=M, aet_bin=L, pr_bin=M, soil_bin=M, ro_bin=L}	0.016

[45] Michael A. Wulder, David P. Roy, Volker C. Radeloff, Thomas R. Loveland, Martha C. Anderson, David M. Johnson, Sean Healey, Zhe Zhu, Theodore A. Scambos, Nima Pahlevan, Matthew Hansen, Noel Gorelick, Christopher J. Crawford, Jeffrey G. Masek, Txomin Hermosilla, Joanne C. White, Alan S. Belward, Crystal Schaaf, Curtis E. Woodcock, Justin L. Huntington, Leo Lymburner, Patrick Hostert, Feng Gao, Alexei Lyapustin, Jean-Francois Pekel, Peter Strobl, and Bruce D. Cook. 2022. Fifty years of Landsat science and impacts. *Remote Sensing of Environment* 280 (Oct. 2022), 113195. doi:10.1016/j.rse.2022.113195

[46] Kornelia Anna Wójcik-Długoborska, Maria Osińska, and Robert Józef Bialik. 2022. The Impact of Glacial Suspension Color on the Relationship Between Its Properties and Marine Water Spectral Reflectance. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 15 (2022), 3258–3268. doi:10.1109/JSTARS.2022.3166398

[47] Hanqiu Xu. 2006. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal of Remote Sensing* 27, 14 (July 2006), 3025–3033. doi:10.1080/01431160600589179

[48] Dai Yamazaki, Mark A. Trigg, and Daiki Ikeshima. 2015. Development of a global 90m water body map using multi-temporal Landsat images. *Remote Sensing of Environment* 171 (Dec. 2015), 337–351. doi:10.1016/j.rse.2015.10.014

[49] Ruyi Yang, Jingyu Hu, Zihao Li, Jianli Mu, Tingzhao Yu, Jiangjiang Xia, Xuhong Li, Aritra Dasgupta, and Haoyi Xiong. 2024. Interpretable machine learning for weather and climate prediction: A review. *Atmospheric Environment* 338 (Dec. 2024), 120797. doi:10.1016/j.atmosenv.2024.120797

[50] Louis Zaugg, Rodolphe Marion, Malik Chami, Xavier Briottet, and Laure Roupioz. 2024. A Physical Method for Optical Characterization of Pollution in Industrial Wastewater Ponds Using Imaging Spectroscopy. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17 (2024), 6029–6044. doi:10.1109/JSTARS.2024.3368750

[51] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833.

A Additional Training Settings

The timeseries of images could have a variable length because of missing images for a given timestep of a minimum quality. We kept only the time series with at least 80% of the timesteps. The input timeseries is zero-imputed to ensure a regular timestep (zero is assigned to no-data on the original image sources). The target is calculated between the two non-imputed time series to avoid introducing additional noise. Since the satellites have different resolutions and quality, we apply random rotation, random translation, and random resize crop to ensure robustness, generalizability, and possible minor misalignment. All input data is standardized. We normalize the target to range -1 to 1. $\alpha = 0.5$ for the loss L_T.

B Climate Subgroups

Table 8 reports the most divergent climate subgroup identified for each task and class. These subgroups correspond to combinations of climate variability levels—discretized into *low (L)*, *medium (M)*, and *high (H)*—across five key variables: precipitation (pr), temperature (tmmx), runoff (ro), evapotranspiration (aet), and soil moisture (soil).

Each subgroup reflects a set of environmental conditions under which the model exhibits the greatest divergence in performance, highlighting systematic weaknesses that recur across samples.

Alongside each subgroup, we report the corresponding Divergence score, which quantifies the severity of the model's underperformance on that specific subgroup. For example, the first group has a F1-score for the NEG class by 0.24 lower than the average. All the divergence scores of the reported subgroups are statistically significant as computed via the Welch-t test as defined in [27].