HydroChronos: Forecasting Long-Term Surface Water Dynamics from Multi-Modal Data

HydroChronos: Forecasting Long-Term Surface Water Dynamics from Multi-Modal Data

DANIELE REGE CAMBRIN*, Politecnico di Torino, Italy and AIKO Space, Italy
ELEONORA POETA, Politecnico di Torino, Italy
ELIANA PASTOR, Politecnico di Torino, Italy
ISAAC CORLEY, Taylor Geospatial, USA
TANIA CERQUITELLI, Politecnico di Torino, Italy
ELENA BARALIS, Politecnico di Torino, Italy
PAOLO GARZA, Politecnico di Torino, Italy

Forecasting surface water dynamics is essential for water resource management and climate change adaptation, but progress is limited by the scarcity of large-scale benchmarks and by the lack of insight into how multi-modal models combine satellite, climate, and topographic signals. In this paper, we provide a detailed characterization of HydroChronos, a large-scale multi-modal spatiotemporal dataset for surface water forecasting that spans over three decades of aligned Landsat 5 and Sentinel-2 imagery, climate variables, and Digital Elevation Models across lakes and rivers in Europe, North America, and South America. We also formalize three forecasting tasks and develop AquaClimaTempo UNet, a spatio-temporal model with a dedicated climate branch and gated fusion mechanism. Across the proposed benchmarks, our model outperforms a persistence baseline by +14% and +11% F1 on change detection and direction-of-change classification, and by 0.1 MAE on magnitude regression. In addition, we present an extended Explainable AI analysis that highlights the most influential climate variables and input channels, offering new insight into the drivers of surface water change.

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

Additional Key Words and Phrases: Spatiotemporal Forecasting, Surface Water Dynamics, Remote Sensing, Explainable AI, Multi-modal Data Fusion

ACM Reference Format:
Daniele Rege Cambrin, Eleonora Poeta, Eliana Pastor, Isaac Corley, Tania Cerquitelli, Elena Baralis, and Paolo Garza. 2026. HydroChronos: Forecasting Long-Term Surface Water Dynamics from Multi-Modal Data. 1, 1 (August 2026), 33 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

The escalating global challenges of water scarcity and climate change, along with their profound impacts on ecosystems and human societies, underscore the importance of understanding and forecasting surface water dynamics [39, 45]. Reliable predictions of future water availability are essential for managing resources in agriculture, energy, and urban systems. As extreme events such

Authors' Contact Information: Daniele Rege Cambrin, Politecnico di Torino, Italy, daniele.regecambrin@polito.it and AIKO Space, Italy, daniele.regecambrin@aikospace.com; Eleonora Poeta, Politecnico di Torino, Italy, eleonora.poeta@polito.it; Eliana Pastor, Politecnico di Torino, Italy, eliana.pastor@polito.it; Isaac Corley, Taylor Geospatial, USA, isaac.corley@taylorgeospatial.org; 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.

ACM XXXX-XXXX/2026/8-ART
https://doi.org/XXXXXXX.XXXXXXX

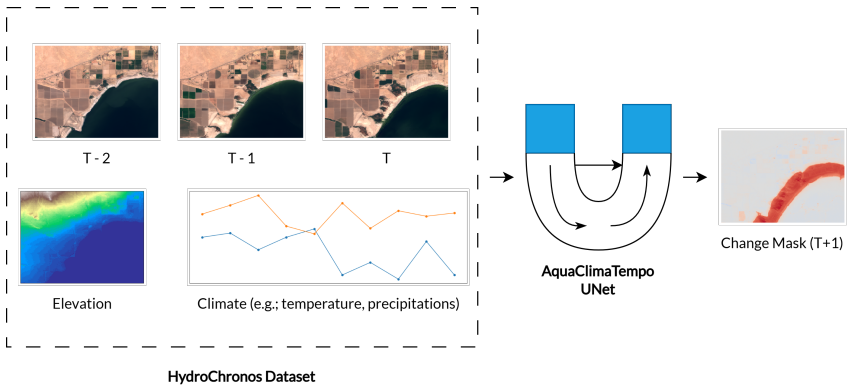


Fig. 1. Forecasting future water landscapes: The HydroChronos dataset (left) provides rich multi-modal inputs including satellite image series, elevation, and climate data. Our *AquaClimaTempo UNet* (center) processes this information to predict future surface water dynamics, generating a *change mask*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

as droughts and floods become more frequent, accurate forecasting models are increasingly critical for enabling timely interventions and improving resilience [13, 20]. However, the inherently complex and spatio-temporal nature of surface water dynamics poses significant challenges, requiring models capable of capturing interactions between hydrological processes and environmental drivers [44].

Despite growing interest in this area, progress is still limited by both data and methodological challenges. On the one hand, large-scale, multi-modal datasets tailored for forecasting tasks remain scarce, with satellite observations often fragmented or poorly integrated with climate and topographic information. On the other hand, beyond data availability, there is still limited understanding of how spatio-temporal models effectively leverage heterogeneous inputs (e.g; satellite imagery, climate variables, and elevation) and how reliable and interpretable their predictions are in practice.

In this work, we provide a comprehensive analysis of the HydroChronos benchmark and the AquaClimaTempo UNet (ACTU) forecasting framework previously introduced in [34]. We present additional details on the construction, characteristics, and limitations of the dataset, offering a deeper understanding of its spatial, temporal, and multi-modal components. We further investigate key design choices of ACTU, including loss formulations, training strategies, and architectural components such as the gating mechanism for multi-modal fusion.

Beyond predictive performance, we conduct a comprehensive Explainable AI (XAI) analysis to better understand model behavior and decision-making processes. This analysis provides insights into how different input modalities contribute to predictions, identifies the primary drivers of surface water change, and highlights the strengths and limitations of the forecasting framework. Together, these investigations provide a more complete understanding of both the benchmark and the modeling approach, offering practical guidance for future research on surface water forecasting.

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

- We provide an extended characterization of HydroChronos, including its construction pipeline, regional composition, temporal coverage, and multimodal structure for water dynamics forecasting.
- We formalize three surface water dynamics forecasting tasks to establish a standardized benchmark for spatio-temporal predictive modeling.

- We develop ACTU as a multimodal baseline that combines satellite imagery, climate variables, and DEM information, and we analyze its architecture, training strategy, loss functions, and gating mechanism in detail.
- We perform a comprehensive XAI study to understand how climatic, temporal, and spatial factors shape predictions and to identify model strengths and limitations. We provide additional analysis of feature-value contributions at both local and global levels, climate-related subgroups, divergence landscapes and feature-frequency patterns, and per-channel saliency investigations.

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

The remainder of this paper is organized as follows: Section 2 reviews related work; Sections 3 to 5 describe the dataset, tasks, and methodology; Section 6 presents experimental results; Section 7 provides the XAI analysis; and Sections 8 and 9 discuss implications and limitations.

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 the 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 [50] and the Sentinel [14] 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) [27] and the Modified NDWI (MNDWI) [52], leveraging water's spectral reflectance. Machine learning classifiers have also been widely employed for more accurate and robust water body extraction [8, 15]. These efforts have culminated in the development of several large-scale and global surface water datasets [32, 53].

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 [25]. Both traditional machine learning approaches [43] and deep learning-based (e.g, based on Long Short-Term Memory [19]) have been applied in Earth observation data [35, 38]. 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 [37] 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 [47]. Our AquaClimaTempo UNet (ACTU), building on similar architectures[4, 21, 42], 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 autoregressive image forecasting.

Our previous work [34] presents a preliminary multi-modal dataset and a first spatio-temporal model for surface water forecasting. However, it provided only a high-level dataset description, a concise architecture presentation, and limited interpretability analysis. The present work substantially extends the dataset characterization, provides an in-depth analysis of the ACTU architecture and loss formulation, and a comprehensive XAI study with subgroup discovery and Shapley-based feature attribution.

2.3 Explainable AI in Earth Sciences

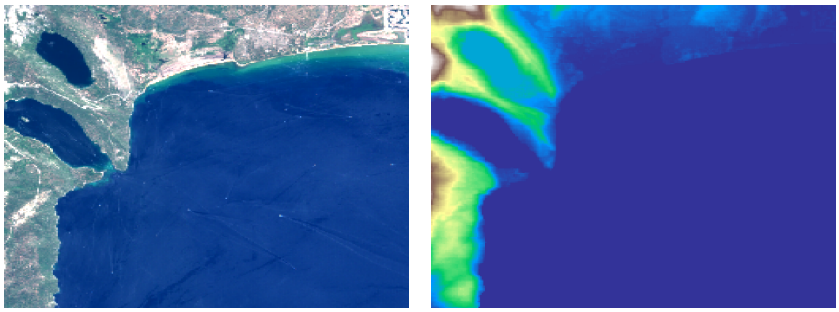
The complexity of AI models in critical domains like Earth sciences demands transparency and interpretability [2, 48]. XAI techniques have been applied to environmental models to identify key input features or understand model behavior [54]. In our case, the integration of climate variables such as precipitation, temperature, and evapotranspiration is well-established as essential for accurate hydrological modeling [9, 17]. 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

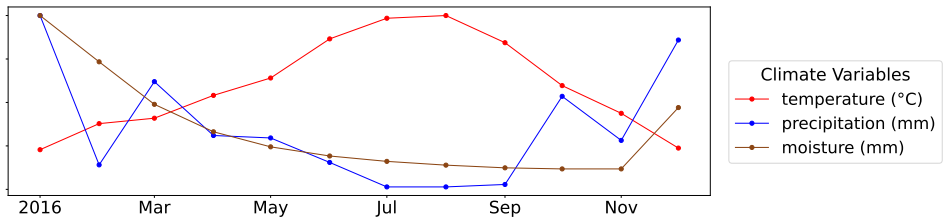
In this section, we present the newly created dataset: HydroChronos. The dataset is composed of time series of images from Landsat-5 and Sentinel-2, time series of climate variables from TERRACLIMATE [1], and a DEM. The selected lakes' and rivers' names are derived from HydroLAKES [28] and HydroRIVERS [22] and cover USA, Europe, and Brazil as shown in Figure 3. A representative sample of the dataset sources is shown in Figure 2.

Table 1 reports the composition of HydroChronos across the three regions. Europe and the United States contribute both lake and river patches, while Brazil exclusively provides river patches, reflecting its role as the fine-tuning split (see Section 3.4). The US and Europe are selected due to the larger availability of samples from Landsat-5 (NASA mission) and Sentinel-2 (ESA mission), while Brazil provides a large source of samples due to its high surface water share (4.1% of the global permanent water, 47% of South America) and large fluctuations in permanent water [16, 32]. The larger number of Brazilian river patches relative to European ones reflects the denser river network in tropical regions and the finer river segmentation at lower Strahler orders.

Candidate patches were generated by tiling each region with a fixed grid and spatially joining against HydroLAKES (lakes ≥ 100 km^2) and HydroRIVERS (flow order ≤ 4). Each patch was then assigned a water coverage score derived from the JRC Global Surface Water maximum extent layer [32], and only patches within a target water coverage range were retained. Overlapping patches



(a) RGB features associated with the sample (b) DEM feature associated with the sample



(c) Climate time series associated with the sample

Fig. 2. A sample in its three modalities. Each sample, collected at a given timestamp, contains optical features (RGB channels only for visualization) at that timestamp, DEM features, and the associated climate time series. Temperature is expressed in °C, precipitation in mm, and moisture in mm.

Table 1. HydroChronos composition by region. Each patch is a fixed spatial sample extracted around a water body. Brazil contains no lake samples by design, as the region was selected to represent river-dominated tropical hydrology.

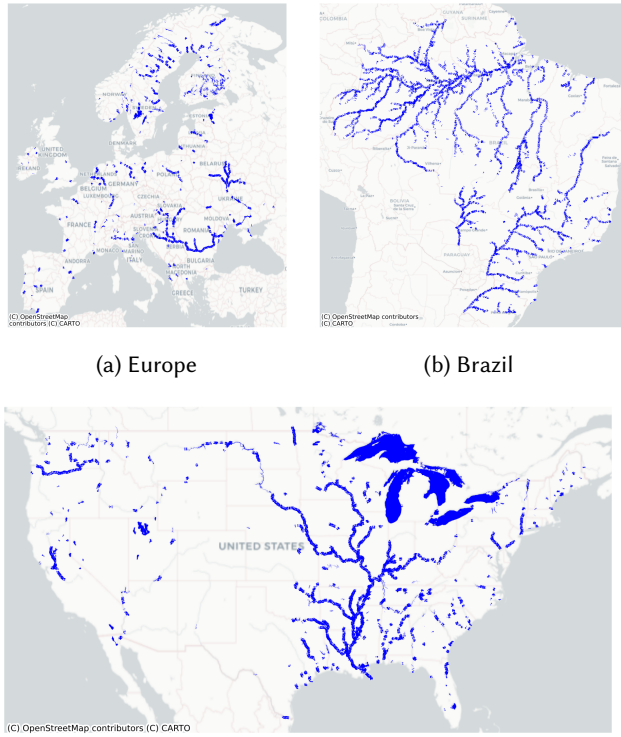
Number of	Europe	US	Brazil	Total
Lakes	78	103	0	181
Rivers	57	38	57	152
Lake Patches	855	2,073	0	2,928
River Patches	762	2,655	9,614	13,031

were resolved by preferring those with mixed land-water content (water coverage closest to 50%), implicitly favoring areas with dynamic rather than permanently flooded water boundaries.

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



(a) Europe (b) Brazil

(c) United States

Fig. 3. Distribution of lakes and rivers in HydroChronos

conditions and minimize unrelated seasonal variability, we selected May-August images (Northern Hemisphere summer).

Table 2. 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

Recognizing the spectral differences between the two sensors, we selected a consistent set of spectral bands. Sentinel-2 provides 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 2. All imagery is provided at a spatial resolution of 30m and transformed to the WGS84 coordinate reference system. An RGB version of a sample can be seen in Figure 2a.

Table 3. Per-region image statistics for HydroChronos. *Median TS* is the median number of yearly observations per time series. Image counts refer to individual patches (one image per patch per year).

	Europe	US	Brazil
Landsat Median TS	21	21	20
Sentinel Median TS	10	9	9
N Landsat Images	19,884	63,885	58,887
N Sentinel Images	5,709	13,139	61,758

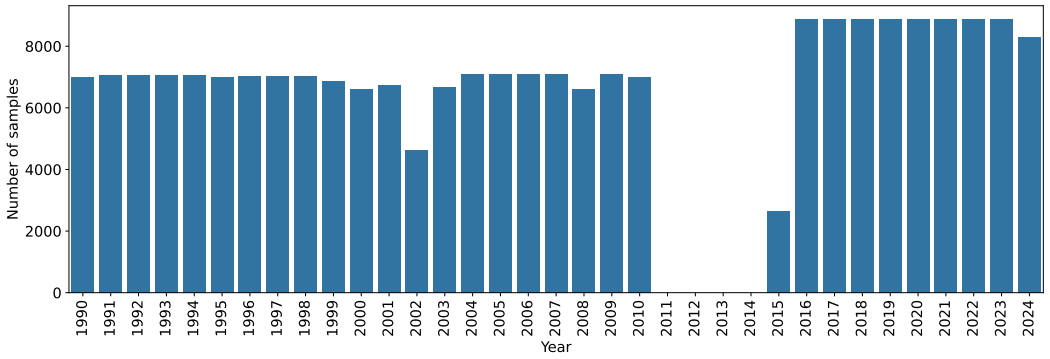


Fig. 4. Number of images per year in HydroChronos. The gap between 2010 and 2015 corresponds to the deliberate exclusion of Landsat-7 acquisitions.

In Figure 4, we report the number of images per year for each sensor. The distribution is relatively uniform across the Landsat period (1990–2010), with minor dips in years affected by higher cloud coverage or reduced acquisition frequency. The Sentinel-2 series starts in 2015 with fewer samples as the satellite reaches full operational capacity progressively and ends in 2024. The 2011-2014 gap between the two sensors is a deliberate design choice: years covered exclusively by Landsat-7, whose Scan Line Corrector failure (2003) introduces systematic striping artifacts, were excluded to preserve image quality. Table 3 complements this view with per-region statistics: median time series lengths are consistent across regions (21 years for Landsat, 9-10 for Sentinel-2), while image counts reflect the different numbers of patches per region reported in Table 1.

Cloud masks are provided along the dataset from the quality mask of the original Landsat-5 and Sentinel-2 products.

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 GLO-30 dataset [3], which offers near-global coverage at approximately 30m spatial resolution, consistent with the imagery stack. A representative sample is shown in Figure 2b. This single-timestep layer captures the terrain elevation for each study area, supporting tasks such as watershed delineation, flow accumulation analysis, and characterization of the topographical influence on water body dynamics [29]. Although we include only the raw elevation raster, first- and second-order terrain derivatives, such as slope, aspect, plan and profile curvature, and the Topographic Wetness Index (TWI), can be computed directly from the GLO-30 layer using standard GIS or remote sensing toolkits.

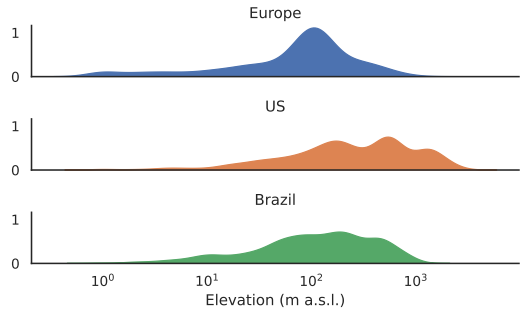


Fig. 5. KDE plot of mean patch elevations per region. Water-surface pixels (elevation ≤ 0m) are excluded.

The three study regions span markedly different elevation distributions, as illustrated in Figure 5. European patches show a compact, unimodal distribution centered near 100 m, consistent with a mix of lowland floodplain rivers and perialpine lakes. US patches exhibit a pronounced multimodal character, with one mode at 50-100 m (Great Lakes basin and eastern lowland sites) and a second at 300-800 m (Rocky Mountain and western montane sites), reflecting the high diversity of the region. Brazilian patches form a broad, right-skewed distribution centered around 100-400m, consistent with the Cerrado plateau above the Amazonian lowlands.

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 as shown in Table 4. 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 2c. These climate variables can be used to analyze the relationship between climatic conditions and observed changes

Table 4. TERRACLIMATE variables included in HydroChronos with their abbreviations and units.

Variable	Abbrev.	Unit
Actual evapotranspiration	AET	mm
Climate water deficit	DEF	mm
Reference evapotranspiration	PET	mm
Precipitation accumulation	PPT	mm
Runoff	Q	mm
Soil moisture	SOIL	mm
Downward shortwave radiation	SRAD	W/m^2
Snow water equivalent	SWE	mm
Maximum temperature	T_{max}	°C
Minimum temperature	T_{min}	°C
Vapor pressure	VP	kPa
Vapor pressure deficit	VPD	kPa
Palmer Drought Severity Index	PDSI	–
Wind speed (10m)	WS	m/s

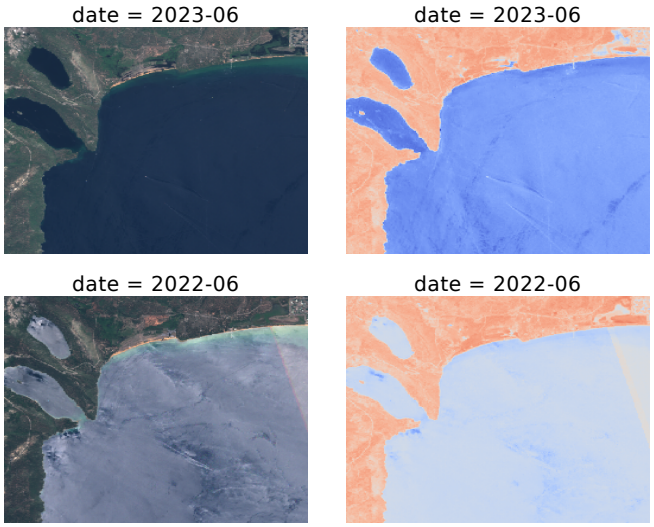


Fig. 6. RGB sample at two different timesteps (on the left) and the corresponding MNDWIs (on the right). MNDWI ranges from -1 (red) to 1 (dark blue).

in water bodies captured by the satellite imagery. Since TERRACLIMATE operates at a coarser spatial resolution than imagery, climate forcing is best interpreted as a catchment-scale driver rather than a pixel-level signal. We reduce each climate patch to a single mean value per variable per month.

The monthly temporal resolution of TERRACLIMATE is finer than the yearly satellite acquisitions, providing up to 12 climate observations per imagery timestep and enabling the model to capture intra-annual climate patterns such as drought onset or seasonal precipitation anomalies.

3.4 Splits

Since each water basin exhibits distinctive hydrological behavior that is difficult to characterize through simple static variables, even between geographically proximate bodies, we adopt a temporal generalization strategy rather than a spatial one.

The Landsat-5 subset (1990–2010) constitutes the pretraining split. Despite its lower image quality and higher noise levels relative to Sentinel-2, its 21-year temporal span provides a rich signal of long-term hydrological dynamics across diverse climatic zones. The Sentinel-2 subset (2015–2024) is divided geographically: Brazilian river patches serve as the fine-tuning split, used to align features learned from Landsat-5 to the Sentinel-2 sensor characteristics and spectral response; European and US patches constitute the test split, used for final evaluation.

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 7.

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: $MNDWI = (G - SWIR)/(G + SWIR)$. The values go from -1 to 1, where values near 1 tend to be water. This approach, while prone to noise, has the advantage of predicting directly the changes regarding physical properties (e.g., icing [46], turbidity [51], pollution [23, 24, 55]) rather than only focusing on water extents and avoiding a costly annotation process. An example can be seen in Figure 6, 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 time series P and a future time series F (which is directly temporarly following P), both composed of MNDWIs of shape $W \times H$, our target change map D (shape $H \times W$) is defined as: $D = median(P) - median(F)$, 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. MNDWI $\in [-1, 1]$, $D \in [-2, 2]$. Based on the analysis of the data, the target distribution is strongly concentrated near zero (the median is 0.01, and the 75th percentile is 0.06), as expected.

The past and future time series P and F are constructed via a sliding window over each patch's full time series. For pretraining on Landsat-5, we use non-overlapping windows of length 5 years each: given a 20-year series from 1990 to 2009, this yields three input-output pairs (P: 1990-1994 $\rightarrow$ F: 1995–1999; P: 1995–1999 $\rightarrow$ F: 2000–2004; P: 2000-2004 $\rightarrow$ F: 2005-2009), substantially augmenting the number of training samples relative to the raw patch count. For fine-tuning and testing on Sentinel-2, the shorter 10-year span (2015–2024) yields a single pair per patch (P: 2015–2019 $\rightarrow$ F: 2020–2024). Consequently, the effective number of time series used for training is approximately three times the number of Landsat patches, while fine-tuning and test counts match the patch counts in Table 1 directly.

4.2 Change Detection

The first proposed task is binary change detection, which can be framed as binary semantic segmentation. Given a time series P (as defined in Section 4.1), a target D (as defined in Section 4.1), and a threshold t to define what we consider a relevant change, we create a binary mask $M_c = |D| > t$. The task focuses on creating a model to predict M_c.

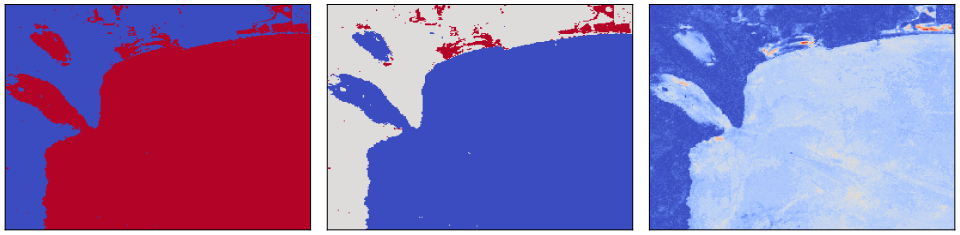
4.3 Direction Classification

This task can be framed as a multiclass semantic segmentation task with 3 classes: negative, positive, or no change. Given a time series P (as defined in Section 4.1), a target D (as defined in Section 4.1), 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 time series P (as defined in Section 4.1) and a target D (as defined in Section 4.1), the task focuses on creating a model to regress the values of $|D|$. This task can be framed as pixel-wise regression. Preliminary experiments also

tried to address the regression of D, but with little success, so we reported only these settings as baseline, leaving this last task for future work.



(a) Change Detection (b) Direction Classification (c) Magnitude Regression

Fig. 7. Visual example of tasks for Lake Tahoe. 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). In magnitude regression, the values range from 0 to 2 (blue to red).

5 Methodology

In this section, we describe ACTU (AquaClimaTempo UNet), our proposed baseline for multimodal spatio-temporal water change prediction. The architecture is designed to jointly process three heterogeneous input modalities (satellite image time series, a static DEM, and monthly climate variable time series) and produce a pixel-wise prediction map. We then describe the regression loss employed for the magnitude estimation task.

5.1 AquaClimaTempo UNet

Throughout this section, T denotes the number of yearly image observations (set to 5), C the number of spectral bands (6), $W \times H$ the spatial dimensions of each patch (256×256), T_1 the number of monthly climate timesteps per image (12), C_1 the number of climate variables (5), and L the number of pyramid levels from the backbone (4 for the ConvNeXtV2). The AquaClimaTempo UNet (ACTU) architecture is depicted in Figure 8.

If a DEM of shape $1 \times 1 \times W \times H$ is provided, it is repeated T times and concatenated along the channel axis with the image time series before being passed to the feature extractor, yielding an input of shape $T \times (C + 1) \times W \times H$.

The image time series (shape $T \times C \times W \times H$, or $T \times (C + 1) \times W \times H$ with DEM) is processed by the *Pyramid Image Feature Extractor* (e.g, ConvNext [49]). 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. Crucially, fusion is performed before temporal aggregation, allowing the network to weight climate contributions per-timestep based on the corresponding optical context.

The *ConvLSTM* layers temporally aggregate the T fused representations at each pyramid level, producing a representation for the time series for each of the L levels of shape from F_0 to F_L.

Each level output F_l serves as the skip connection for the corresponding stage of the UNet decoder's expanding path, replacing standard single-timestep encoder skip connections with

temporally-integrated representations. This allows the decoder to leverage full historical context at each spatial scale when producing the final prediction mask of shape $1 \times W \times H$.

ConvLSTM [41] is chosen over purely attention-based temporal encoders to preserve spatial structure while capturing sequential dependencies, at a lower computational cost and for short sequences.

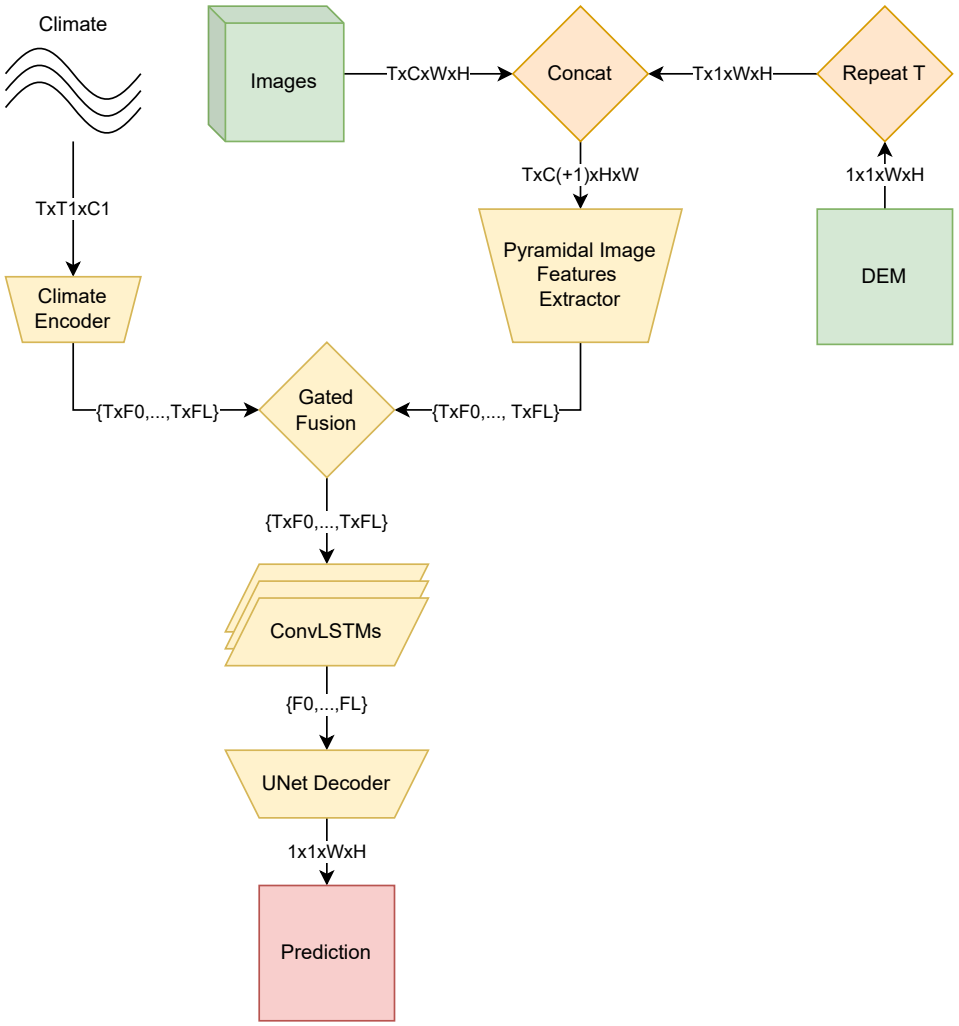


Fig. 8. AquaClimaTempo UNet (ACTU) architecture. If DEM is provided, it is repeated once per sample in the image time series and concatenated along the channel axis. The *Pyramidal Image Feature Extractor* provides multiscale embeddings. If a climate time series is provided, the *climate encoder* provides multiscale embeddings which are *gate fused* with the image embeddings. *ConvLSTMs* provide multiscale embeddings for the time series, which are used in the *UNet decoder* to provide the final prediction.

5.1.1 Climate Encoder. The climate encoder processes X_{clim}, one time series 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. Each is first summarized by a single-layer LSTM Φ_{LSTM} applied along the T_1 temporal axis, whose final hidden state captures the climate history preceding each yearly image. The hidden state is then 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)).

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

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

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

The number of upsampling steps S at each level is chosen such that the output resolution of K_l matches the spatial dimensions of the corresponding image feature I_l, enabling element-wise fusion in the gated fusion module.

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:

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

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

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

Each F_l is then used in the decoder to make the final prediction. It is important to note that α is computed spatially and channel-wise, allowing the network to selectively suppress climate influence in regions where the optical signal is already informative (e.g., cloud-free open water) while amplifying it where the optical signal is ambiguous (e.g., cloud-masked or mixed pixels).

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 HuberLoss 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 $S = \{s_1, ..., s_M\}$ are M different scale factors and D_x is the bilinear downscale operation with factor x. The first term $L(P,T)$ corresponds to the full-resolution loss.

5.2.2 Wavelet Loss. The Discrete Wavelet Transform (DWT) provides a multi-resolution decomposition of a 2D signal into spatially localized frequency sub-bands [26]. Unlike the Fourier transform, which decomposes a signal into global sinusoids, wavelets retain spatial locality, making them suitable for images where changes are geographically concentrated (e.g., shoreline shifts, localized drought effects). The DWT applies a pair of low-pass and high-pass filters recursively: the low-pass filter produces an approximation Y_L capturing coarse structure, while the high-pass filter produces detail coefficients Y_H at three orientations (horizontal, vertical, and diagonal) per decomposition level. We use the Daubechies-2 (DB2) wavelet [12] because it produces smoother approximation coefficients than the simpler Haar wavelet, but it has a sufficiently short support to see local changes and is more efficient than Symlets, Coiflets, and Biorthogonal.

The DWT is applied to both prediction and ground truth. 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: $L_T = \lambda L_{MS} + (1 - \lambda)L_W$. The multiscale component L_{MS} encourages spatially coherent predictions at coarse resolutions, reducing the influence of fine-grained noise on the overall loss magnitude. The wavelet component L_W complements this by penalizing errors in oriented high-frequency structures (shoreline edges, channel boundaries) that would be suppressed by downscaling alone.

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. While evaluation metrics computed over the entire test set provide an overall estimate of predictive quality, they can obscure systematic patterns of failure concentrated within specific climate regimes or environmental contexts. Similarly, aggregate evaluation may hide conditions under which the model performs substantially better than average. Identifying these localized behaviors is important both for understanding model limitations, robustness, and potentially unsuitable deployment conditions, and for characterizing the environmental settings under which the model can be considered more reliable and trustworthy. To uncover and characterize such predictive behaviors, we adopt a post hoc analysis framework based on subgroup discovery and feature attribution. Specifically, we aim to: (i) identify interpretable subsets of samples associated with significantly different predictive performance, and (ii) explain which environmental and climatic factors contribute to these localized variations in model behavior.

Climate Subgroup Discovery. Environmental and climatic factors can influence both the underlying data distribution and the behavior of machine learning models. We investigate whether specific combinations of these factors are systematically associated with variations in predictive quality. To this end, we employ subgroup discovery to identify interpretable climatic regimes characterized by significantly different model performance. Formally, 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 environmental and climate descriptors $A(x_i) = \{a_1, \ldots, a_k\}$ derived from climatic and geographic variables. Since many environmental variables are continuous, we discretize them into semantically meaningful intervals through feature binning (e.g., low precipitation, high temperature). Discretization serves two main purposes. First, it enables the identification of groups of samples sharing similar environmental characteristics, rather than focusing on isolated numerical values that may not generalize. Second, it improves interpretability by producing human-readable climate profiles that can be directly analyzed by domain experts.

Let $J = \{(a_1 = v_1), \ldots, (a_l = v_l)\}$ be a set of attribute-value pairs, where each pair $(a_j = v_j)$ specifies that environmental attribute a_j assumes the discretized value or interval v_j. The *subgroup* associated with J is the subset of samples in $\mathcal{D}$ satisfying all conditions in J:

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

For example, the subgroup $J = \{(\texttt{temperature = high}), (\texttt{precipitation = low})\}$ identifies the subset of samples characterized by high temperature and low precipitation conditions. Each subgroup, therefore, represents an interpretable environmental or climatic regime characterized by a coherent combination of conditions.

The objective of subgroup discovery is to identify subgroups whose predictive performance significantly differs from the overall dataset performance. To quantify this difference, we adopt the notion of subgroup performance divergence [30]. Given a scalar performance metric $m : \mathcal{D} \to \mathbb{R}$, the divergence of a subgroup S is defined as:

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

where $m(S)$ denotes the metric evaluated over subgroup S, while $m(\mathcal{D})$ represents the corresponding global performance computed over the entire dataset.

The magnitude of $\Delta_m(S)$ quantifies how strongly the subgroup behavior deviates from the global average. Depending on the selected metric m (e.g., precision, recall, F1-score, false negative rate), positive or negative divergence values identify environmental conditions associated with unusually strong or weak predictive behavior. For example, a large negative divergence in F1-score indicates that the model performs substantially worse than average within that specific subgroup.

To automatically identify such patterns, we employ `DivExplorer` [30], a divergence-based subgroup discovery framework grounded in frequent pattern mining. The approach explores the combinatorial space of attribute-value conditions to extract statistically meaningful and sufficiently representative subgroups. More specifically, it searches for combinations of categorical or discretized environmental attributes that frequently co-occur within the dataset, computes their performance, and evaluates whether the predictive performance associated with these subsets significantly deviates from the global behavior.

Candidate subgroups are generated by identifying frequent combinations of attribute-value conditions satisfying a minimum support threshold θ. The support of a subgroup corresponds to the proportion (or number) of samples in the dataset satisfying the subgroup conditions. Introducing a minimum support constraint is important to avoid extracting highly specific patterns associated with only a few samples, which may correspond to noise, outliers, or statistically unreliable behaviors.

By enforcing a minimum subgroup size, the analysis focuses on subgroups that are sufficiently representative and meaningful within the dataset.

For each candidate subgroup, `DivExplorer` computes the subgroup support, quantifying how representative the subgroup is within the dataset; the subgroup predictive performance according to a target metric; and the divergence score with respect to the global performance. This process enables the identification of environmental and climatic regimes where the model systematically underperforms or overperforms relative to the overall dataset behavior.

The extracted subgroups are represented as interpretable rule-based descriptions (e.g., *high temperature* $\wedge$ *low precipitation seasonality*), making the resulting analysis transparent and interpretable. Subgroup interpretability makes subgroup discovery suitable for auditing complex environmental machine learning systems, identifying climate-dependent biases, and revealing subgroup-specific model limitations or strengths that may remain hidden under aggregate evaluation metrics.

This analysis serves multiple purposes. First, it enables subgroup auditing of model reliability across heterogeneous environmental conditions. Second, it provides interpretable descriptions of failure modes that can support domain experts in understanding climate-dependent biases and limitations of the predictive system. Finally, the discovered subgroups can guide subsequent model refinement, data collection, and bias mitigation strategies by highlighting underrepresented or systematically challenging climate regimes.

Feature Attribution. While subgroup discovery identifies environmental conditions associated with significantly different performance, it does not directly explain which variables contribute to these variations across subgroups. To complement the subgroup analysis, we employ feature attribution methods based on Shapley values [40], which provide a principled way to quantify the contribution of each climate or environmental factor to the divergence score, both locally and globally.

Local contribution. Shapley values originate from cooperative game theory and measure the contribution of an element to a total score by considering its marginal effect across all possible combinations. In our setting, the elements correspond to the attribute-value conditions defining a subgroup, while the target quantity corresponds to the subgroup divergence Δ_m.

Given a subgroup described by $J = \{a_1 = v_1, \ldots, a_l = v_l\}$, the contribution of each attribute-value pair $a = v \in J$ to the subgroup divergence is defined as [31]:

$$\phi(a = v \mid J) = \sum_{Y \subseteq J \setminus \{a=v\}} \frac{|Y|!(|J| - |T| - 1)!}{|J|!} \left[\Delta_m(T \cup \{a = v\}) - \Delta_m(T)\right].$$

Here, Y denotes a subset of attribute-value conditions selected from J excluding condition (a, v). Therefore, Y represents a partial subgroup description, while $Y \cup \{a = v\}$ corresponds to the same subgroup description after adding the condition $a = v$. The term $\Delta_m(Y \cup \{a = v\}) - \Delta_m(Y)$ measures the marginal contribution of condition $a =, v$ when added to the partial subgroup description Y. The weighting factor averages this contribution across all possible orderings of the subgroup conditions.

Intuitively, the local attribution score quantifies how much a specific climate or environmental factor contributes to the divergence observed for a subgroup. Positive values indicate that the condition factors the subgroup divergence, while negative values indicate that it reduces or counteracts it. The magnitude reflects the strength of the contribution.

For example, consider the subgroup $J = \{(\texttt{temperature} = \texttt{high}), (\texttt{precipitation} = \texttt{low})\}$. The corresponding attribution scores quantify whether the observed divergence is mainly associated with high temperature, low precipitation, or their combination, and to what extent each condition contributes to the divergence.

Global contribution. Subgroup-specific contribution explains the contribution of environmental conditions within a specific subgroup, representing a local contribution of the attribute-value pairs in a subgroup to the subgroup itself. We are also interested in identifying which conditions consistently contribute to divergence across the overall subgroup exploration. This score follows the global divergence formulation of [31], adapted to our setting, where attribute-value conditions define climate subgroups and Δ_m is the divergence of the performance metric m.

Given an attribute-value condition $a = v$, we define its global contribution as:

$$\Phi(a = v) = \sum_{J \in \mathcal{J}: a=v \notin J} \frac{|J|!(|A| - |J| - 1)!}{|A|!} \left[\Delta_m(J \cup \{a = v\}) - \Delta_m(J)\right],$$

where $|A|$ denotes the total number of environmental attributes, and $\frac{|J|!(|A|-|J|-1)!}{|A|!}$ is a Shapley weighting term accounting for all possible orderings of the subgroup conditions.

Intuitively, $\Phi(a = v)$ measures the average marginal effect of adding condition $a = v$ across the subgroup descriptions explored during subgroup discovery. 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.

5.3.2 Per-channel Saliency. A central goal in explainable artificial intelligence is to identify which input components most significantly influence a model's predictions. Beyond interpretability, this analysis has practical implications, such as identifying less informative channels that could potentially be removed to reduce computational overhead and data acquisition requirements in our spatiotemporal setting. A common strategy for estimating input relevance is perturbation-based analysis, which involves perturbing the input and measuring the resulting change in the model's output. Perturbation-based techniques [5, 11, 56] provide an intuitive and model-agnostic mechanism for estimating the contribution of input dimensions according to their impact on the model outputs.

In this work, we adopt a perturbation-based strategy to compute *per-channel saliency*, aimed at quantifying the contribution of each input channel to the predictive performance of the model. We adapted the methodology proposed in [33] to the spatiotemporal setting considered in this work.

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 score of channel c for a given sample is defined as the change induced by the perturbation in a target performance metric m:

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

where y denotes the ground-truth target, $\hat{y}$ is the original model output, and $\hat{y}^{(-c)}$ is the output obtained after perturbing channel c. A larger positive value of Δm_c indicates that channel c contributes more strongly to the predictive performance of the model, since removing the channel causes a larger degradation in the evaluation metric. Conversely, values close to zero indicate that the model is relatively insensitive to the information carried by that channel. Negative values may also occur, indicating that perturbing the channel improves the evaluation metric, suggesting that the corresponding information may be noisy, redundant, or detrimental to the model.

To obtain a dataset-level estimate of channel importance, we average the saliency scores across the test set. 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 equally weighted generalized dice and focal losses. The vision backbone for the encoder is ConvNextV2 [49], in base (ACTU) and large (ACTU-L) versions. The time series 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 time series 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. We standardize per-channel the input data.

To avoid overwhelming the model with redundant information and to limit multicollinearity among highly correlated variables (e.g., AET and PET, VP and VPD), we select a subset of five climate variables that jointly cover the primary physical drivers of surface water change: maximum temperature (tmmx), which governs evaporative demand and ice/snow melt; actual evapotranspiration (aet), which quantifies direct water loss from the surface; runoff (ro) and precipitation (pr), which control water inflow; and soil moisture (soil), which captures subsurface storage and lagged hydrological response [6, 18, 36].

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 5 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 5. Change detection results for models optionally using DEM (D) and climate variables (C). ^{a} indicates statistically significant difference ($p < 0.01$) with respect to persistence according to the t-test. ^{b} 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^a	82.78^a	85.66^a	44.87^a	60.65^a	48.79^a
ACTU	N	Y	88.92^a	85.86^a	86.6^a	45.45^a	53.1^a	45.83^a
ACTU	Y	N	**90.57**	82.89^a	85.75^a	45.19^a	61.01^a	**49.38**^{a}
ACTU	Y	Y	90.53^a	81.71^a	85.08^a	43.68^a	62.33^a	48.67^a
ACTU-L	N	N	90.03^b	84.3^b	86.43^b	**45.46**^{b}	57.34^b	47.94^b
ACTU-L	Y	Y	89.2^b	84.95^b	86.37^b	45.03^b	54.39^b	46.49^b

6.4 Direction Classification

Table 6 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,

Table 6. Direction classification results for models optionally using DEM (D) and climate variables (C). [a] indicates statistically significant difference ($p < 0.05$) with respect to persistence according to the t-test. [b] 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[a]	**49.38**[a]	**30.27**[a]	**90.23**[a]	84.58[a]	86.67[a]	27.73[a]	19.84[a]	19.47[a]
ACTU	N	Y	**29.84**[a]	36.53[a]	27.75[a]	88.44	87.5[a]	87.42[a]	26.16[a]	24.12[a]	**21.38**[a]
ACTU	Y	N	28.18[a]	34.43[a]	25.81[a]	87.73[a]	88.65[a]	87.54[a]	27.39[a]	20.25[a]	19.09[a]
ACTU	Y	Y	28.36[a]	37.35[a]	27.28[a]	88.18[a]	88.17[a]	87.6[a]	27.98[a]	22.06[a]	20.77[a]
ACTU-L	N	N	28.42[b]	38.21[b]	27.62[b]	88.5[b]	88.01[b]	87.76[b]	**28.25**[b]	22.04[b]	20.88[b]
ACTU-L	Y	Y	27.52[b]	34.84[b]	25.31[b]	88.15	88.09	87.55	27.13[b]	21.28[b]	19.44[b]

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 7 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.

Table 7. Magnitude regression results for models optionally using DEM (D) and climate variables (C). ^{a} indicates statistically significant difference ($p < 0.05$) with respect to persistence according to the t-test. ^{b} indicates the statistical difference comparing ACTU-L with the same configuration of ACTU.

Model	D	C	MAE	MAE@10	MAE@20	PC	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2
Constant	N	N	.0351	.142	.1038	-	0	0	0	0	0
Persistence	N	N	.1281	.1892	.171	31.81	23.86	**81.77**	34.98	13.38	**75.72**
ACTU	N	N	**.0261**^{a}	.0873^a	.0611^a	46.45^a	**51.97**^{a}	41.61^a	43.04^a	45.27^a	24.48^a
ACTU	N	Y	.0266^a	.0911^a	.0639^a	44.4^a	51.32^a	41.59^a	42.77^a	42.34^a	18.12^a
ACTU	Y	N	.0297^a	.0886^a	.0643^a	44.46^a	49.36^a	45.09^a	43.64^a	40.67^a	28.26^a
ACTU	Y	Y	.0315^a	.088^a	.0634^a	41.41^a	46.92^a	45.18^a	42.67^a	39.57^a	27.91^a
ACTU-L	N	N	.0275^b	**.0843**^{b}	**.0589**^{b}	**46.62**	50^b	**47.19**^{b}	**45.61**^{b}	44.45^b	27.55^b
ACTU-L	Y	Y	.0282^b	.0923^b	.066^b	43.45^b	51.63^b	38.48^b	40.65^b	43.24^b	21.41^b

6.6 Ablation Studies on Regression Loss

To understand the contribution of the proposed composed loss L_T, we report in Table 8 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.

Table 8. Comparison between the wavelet loss (L_W), multiscale 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	P@0.1	R@0.1	F@0.1	P@0.2	R@0.2	F@0.2
L_T	**.0261**	.0873	.0611	**46.45**	**51.97**	41.61	43.04	**45.27**	24.48	**26.85**
L	.029*	.0861*	.06*	42.9*	46.14*	45.65*	42.25	39.89*	24.25	24.61*
L_{MS}	.0291*	**.0839***	**.0585***	43.82*	45.85*	**49.84***	**44.83**	39.38*	**27.2**	26.75*
L_W	.0285*	.1047*	.0765*	42.1*	51.24	25.11*	29.88*	38.51*	16.24*	18.42*

7 XAI Analysis

In this section, we present the insights derived from the XAI analysis on the behavior of ACTU network. We begin studying how predictive performance varies across interpretable climate regimes

Table 9. Discretization thresholds used to define low (L), medium (M), and high (H) variability bins for climate subgroup discovery.

Variable	L–M	M–H	Max
T_{max}	-0.74	0.04	0.82
$Precip.$	-1.31	-0.26	0.79
$Runoff$	-0.27	0.63	1.54
$Soil$	-3.41	-0.93	1.55
AET	-0.61	0.03	0.68

using subgroup discovery and Shapley-based feature attribution. We then quantify the contribution of each spectral channel and DEM through perturbation-based per-channel saliency.

7.1 Climate Subgroups and Feature Attribution

Climate Subgroups. We identify climate subgroups associated with the strongest deviations from the global model behavior 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 subgroup discovery on the five climate variables, discretized based on their intra-series variability. The discretization thresholds used for the subgroup analysis are reported in Table 9. Each subgroup is defined by combinations of discretized variability levels for the five climate variables used by the model: precipitation (pr), temperature (tmmx), runoff (ro), evapotranspiration (aet), and soil moisture (soil).

Table 10 reports the most and least performant subgroups for the change detection and direction classification, evaluated using the F1-score, and regression tasks, evaluated using the MAE. 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 [31].

Overall, the discovered subgroups reveal that model performance is not uniformly distributed across climate regimes. For change detection, the best-performing subgroups (Row 3) are characterized by high precipitation, along with medium runoff and soil moisture, whereas low precipitation, low runoff regimes with medium evapotranspiration, soil moisture, and maximum temperature are associated with poor change-detection performance (Row 4). For direction classification, low-prediction rates are particularly difficult under high evapotranspiration, high precipitation, and low runoff (Row 6), while stable areas are poorly predicted under medium precipitation, high runoff, and high temperature (Row 7). In regression, the largest errors are associated with high evapotranspiration, low precipitation, low soil moisture, and high temperature (Row 12), suggesting that dry, evaporatively demanding regimes remain challenging for magnitude estimation.

We further inspect the samples contained in the worst-performing subgroups to identify recurrent problematic regions. As summarized in Table 11, Great Salt Lake and Utah Lake appear consistently among the hardest cases for change detection, direction classification, and regression. Rainy Lake and Woods Lake are mainly associated with failures on stable or no-change predictions, while Lake Texoma and Pyramid Lake recur in subgroups with poor regression and positive-change performance. These patterns indicate that the discovered climate regimes capture meaningful and geographically coherent failure modes rather than isolated outliers.

Table 10. Most and least performant subgroups across C, D, and R tasks. Positive divergence indicates above-average performance, while negative divergence indicates below-average performance.

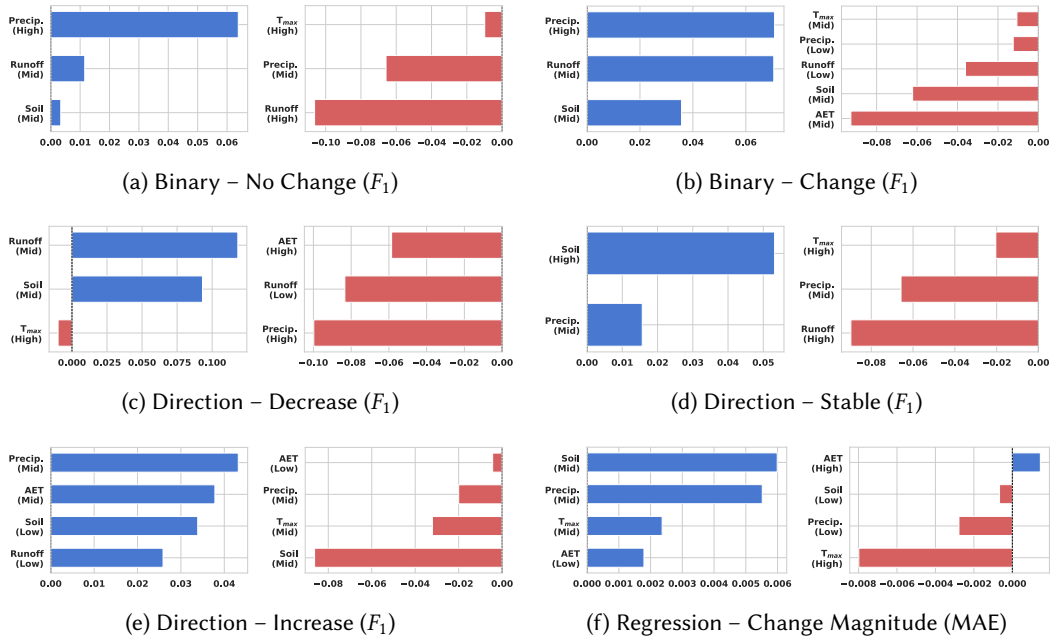
Task	Target/Class	Rank	Subgroup pattern	Score	Div
C	No change (F_1)	Top	Precip.=High, Runoff=Medium, Soil=Medium	0.94	
C	No change (F_1)	Bottom	Precip.=Medium, Runoff=High, Tmax=High	0.68	
C	Change (F_1)	Top	Precip.=High, Runoff=Medium, Soil=Medium	0.67	
C	Change (F_1)	Bottom	AET=Medium, Precip.=Low, Runoff=Low, Soil=Medium, Tmax=Medium	0.28	
D	Decrease (F_1)	Top	Runoff=Medium, Soil=Medium, Tmax=High	0.48	
D	Decrease (F_1)	Bottom	AET=High, Precip.=High, Runoff=Low	0.04	
D	Stable (F_1)	Top	Precip.=Medium, Soil=High	0.94	
D	Stable (F_1)	Bottom	Precip.=Medium, Runoff=High, Tmax=High	0.70	
D	Increase (F_1)	Top	AET=Medium, Precip.=Medium, Runoff=Low, Soil=Low	0.35	
D	Increase (F_1)	Bottom	AET=Low, Precip.=Medium, Soil=Medium, Tmax=Medium	0.07	
R	Change Magnitude (MAE)	Top	AET=Low, Precip.=Medium, Soil=Medium, Tmax=Medium	0.04	
R	Change Magnitude (MAE)	Bottom	AET=High, Precip.=Low, Soil=Low, Tmax=High	0.02	

Table 11. 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	Change	Increase	Change Magnitude
Utah Lake	Change	Increase	Change Magnitude
Rainy Lake	No Change	Stable	—
Woods Lake	No Change	Stable	—
Lake Texoma	—	Increase	Change Magnitude
Pyramid Lake	—	Increase	Change Magnitude

Feature Attribution. We compute the local and global feature-value contributions via 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.

Local subgroup contributions. Figure 9 reports local contribution via Shapley values for the best- and worst-performing subgroups across tasks. These explanations clarify which climate conditions contribute most to each subgroup's divergence. In the binary task, high precipitation supports strong performance for both no-change (9a, left) and when coupled with medium runoff for change classes (9b, left), whereas high runoff mostly contributes to lower-than-average performance to no change detection (9a, right), while medium AET for change one (9b, right). In the direction task, high precipitation is the term contributing most to the subgroup with lower-than-average predictions for the decrease value (9c, right), whereas mid soil contributes most to the subgroup with lower-than-average predictions for the increase value (9e, right). For regression (9f), favorable subgroups are mainly explained by soil moisture and medium precipitation, whereas high evapotranspiration, low precipitation, and high temperature contribute to larger errors.



(a) Binary – No Change (F_1)

(b) Binary – Change (F_1)

(c) Direction – Decrease (F_1)

(d) Direction – Stable (F_1)

(e) Direction – Increase (F_1)

(f) Regression – Change Magnitude (MAE)

Fig. 9. Local Shapley explanations for the highest- and lowest-performing subgroups across all tasks. Blue bars indicate positive feature contributions, while red bars indicate negative contributions. Each panel compares the top subgroup (left) and bottom subgroup (right) according to divergence from the global average metric.

Global Shapley explanations. While the local analysis quantifies how each condition contributes to the divergence of a specific subgroup, it requires examining one subgroup at a time and therefore provides only a local view of model behavior. To obtain a broader picture of the climate factors that systematically affect performance, we can use the Global Shapley values over all discovered subgroups.

Figure 10 and Table 10 summarize the most influential feature–bin combinations. Across tasks, runoff and evapotranspiration emerge as recurrent drivers of performance variation. High runoff is often associated with reduced performance, especially for stable and increasing classes, while high evapotranspiration negatively affects several direction and regression metrics. Conversely, moderate or low precipitation and runoff regimes frequently correspond to improved predictive behavior. These results suggest that the model is more reliable under moderate hydrological variability, while extreme or internally inconsistent climate regimes expose systematic weaknesses.

Divergence distribution analysis. To better characterize the variability of subgroup behaviors, Figure 11 reports the distribution of subgroup divergence scores across all evaluated metrics and classes for the binary, direction, and regression models, considering F1-score, precision, and recall for the classification tasks, and MAE for regression. The violin plots reveal that divergence distributions are strongly asymmetric across tasks. In the binary task (11a), recall-related metrics exhibit the largest variance, especially for the change class, indicating that certain climate regimes strongly amplify or suppress the model's ability to detect changes. In the direction task (11b), the decrease and increase classes show substantially broader divergence distributions than the stable class, confirming that directional prediction is considerably more sensitive to climate variability. For regression (11c), the divergence distribution is comparatively narrow and centered near zero,

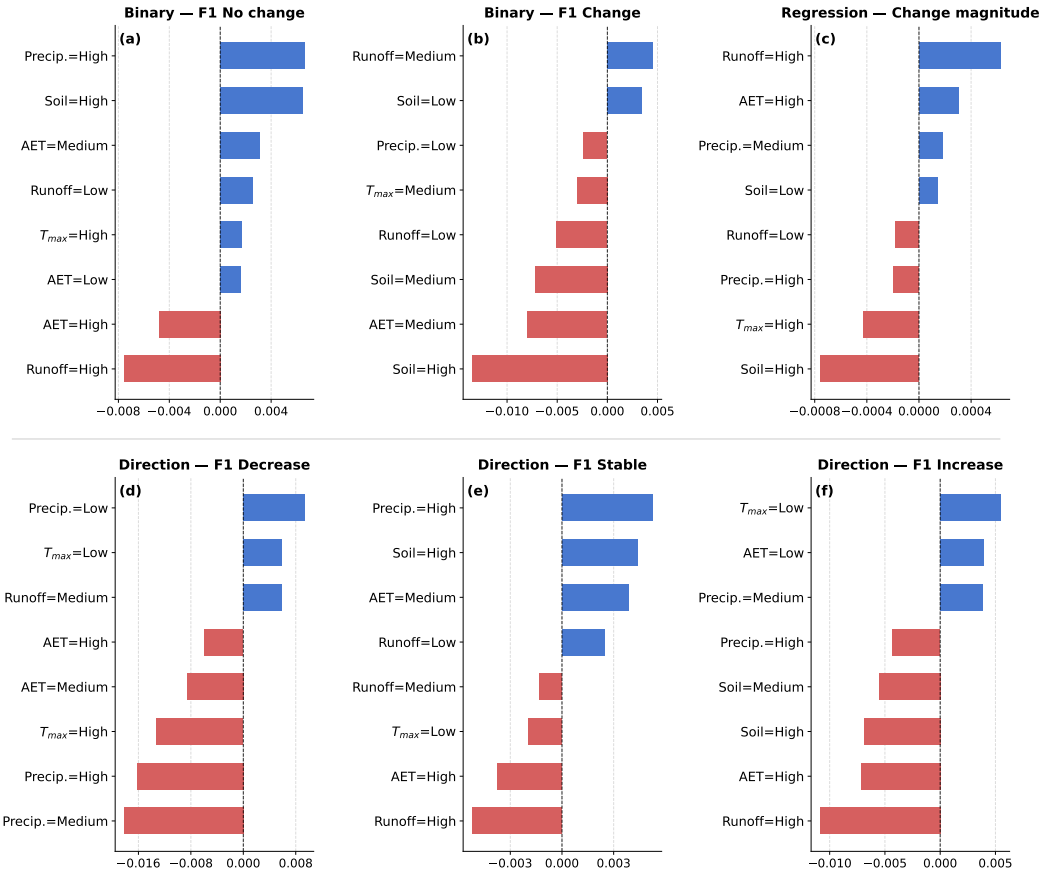
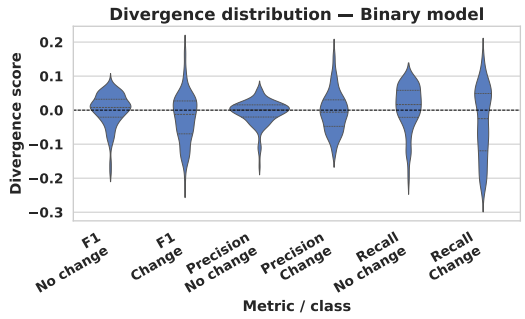


Fig. 10. Global Shapley explanations aggregated across all discovered subgroups for the binary, regression, and direction prediction tasks. Each panel highlights the most influential feature–bin combinations contributing positively (blue) or negatively (red) to model performance. Panels (a)–(b) show binary classification performance for the Non-water and Water classes, respectively; panel (c) shows regression performance for change magnitude prediction (MAE sign-flipped so positive values indicate improvement); and panels (d)–(f) show direction classification performance for the Decrease, Stable, and Increase classes. Across tasks, high runoff and high AET conditions are consistently associated with reduced performance, whereas moderate or low precipitation/runoff regimes frequently correspond to improved predictive behavior.

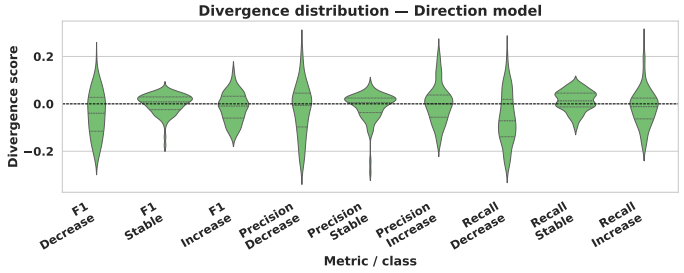
suggesting more stable behavior across climate conditions despite a few strongly challenging regimes. Overall, these distributions confirm that the model errors are not uniformly distributed but concentrated within specific environmental settings.

7.2 Per-channel Saliency

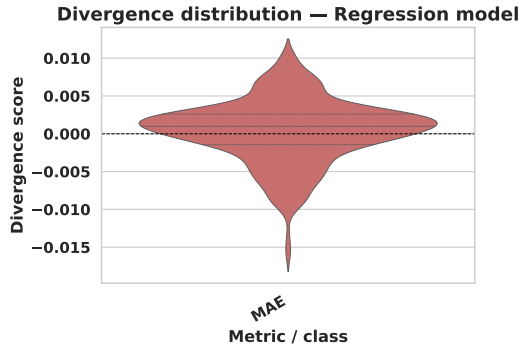
We perform a per-channel saliency analysis to evaluate the contribution of each input modality. We perform perturbation-based per-channel saliency to quantify the importance of each input channel. For classification tasks, saliency is measured as the drop in F1-score when a channel is ablated, while for regression it is measured through the change in magnitude prediction performance. Figure 12 reports the mean contribution of each spectral band and DEM across tasks.



(a) Binary model



(b) Direction model



(c) Regression model

Fig. 11. Distribution of subgroup divergence scores across all evaluated metrics and classes for the binary, direction, and regression models. Positive divergence values indicate subgroups performing above the global average, while negative values indicate below-average performance. For regression, MAE divergence is sign-flipped so that positive values consistently correspond to better performance.

The results reveal a strong task- and class-dependent structure. In the regression task, NIR provides the largest contribution, followed by SWIR1 and SWIR2, confirming the importance of infrared information for estimating water-related magnitude changes. In the binary task, the Change class relies heavily on NIR and SWIR bands, whereas the No Change class shows weaker and more distributed dependencies across RGB channels. Direction classification exhibits the strongest class-specific behavior: NIR strongly supports decreasing-change detection but negatively affects increasing changes, while SWIR channels contribute positively to stable and increasing classes.

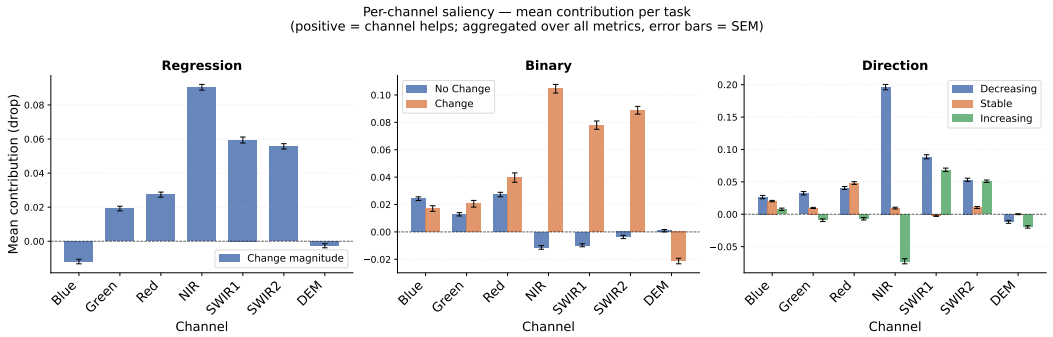


Fig. 12. Per-channel saliency analysis across regression, binary, and direction tasks. Bars show the mean performance drop after perturbing each input channel, with error bars indicating the standard error of the mean (SEM). NIR and SWIR bands make the greatest contributions, particularly for regression and change detection, whereas direction classification shows strong class-specific dependencies. In particular, NIR strongly supports detecting decreases but negatively affects detecting increases. DEM contributes weakly or negatively across most tasks, suggesting limited complementary information under the current setting.

DEM contributes weakly or negatively across most tasks, suggesting that static elevation provides limited complementary information once the spectral time series is available. Overall, the saliency analysis confirms that infrared channels are central for dynamic water-change prediction, while RGB channels mainly provide contextual information for stable-region discrimination.

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.

7.3 Gated fusion analysis

To investigate whether the auxiliary modalities are actively leveraged by the model, we analyzed the distribution of gate activation values α across the full test set. Across all four levels, α follows a sharp unimodal distribution tightly centered around 0.5 with small standard deviations ($\mu \in [0.499, 0.503]$, $\sigma \in [0.022, 0.055]$), with no evidence of bimodality or systematic deviation toward either modality as shown in Figure 13. This indicates that the gated fusion has converged to a near-uniform weighting, functionally equivalent to simple feature averaging, rather than learning spatially adaptive decisions. Combined with the inconclusive modality ablation results, this finding strongly suggests that the auxiliary inputs do not carry sufficient complementary signal to drive meaningful gate adaptation under the current training regime.

8 Discussion and Implications

The results presented in this work collectively advance our understanding of both the capability and the internal mechanics of spatio-temporal models for surface water dynamics forecasting.

Performance gains and multi-modal integration. ACTU consistently and significantly outperforms both the constant and persistence baselines across all three tasks, confirming that learned spatio-temporal representations from historical satellite imagery can meaningfully anticipate future surface water dynamics. The gated fusion analysis reveals that the model converges to near-uniform weighting across all pyramid levels, suggesting that the auxiliary inputs do not provide sufficient

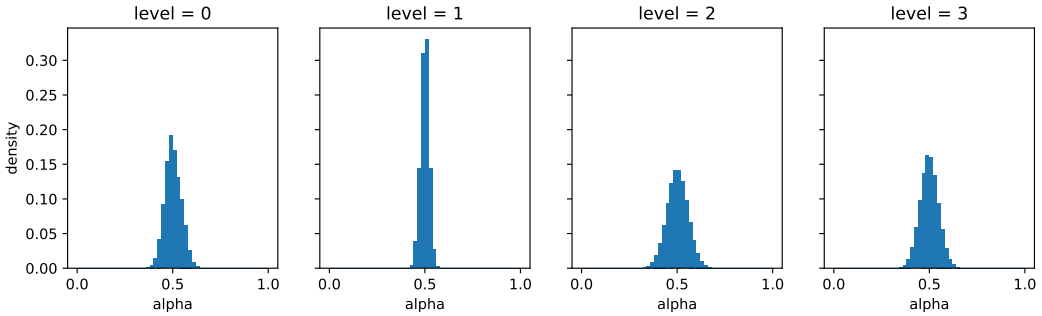


Fig. 13. Gating α values at different levels of the pyramidal resolution.

complementary signal under the current data regime. We identify two candidate explanations: the DEM is temporally misaligned in most of the cases and co-varies weakly with the optical time series, and many TERRACLIMATE variables (e.g., runoff and soil moisture) are outputs of linear water balance models, known to underperform in the non-linear hydrological regimes that our subgroup analysis identifies as the hardest cases.

Spectral band importance and task dependency. The per-channel saliency analysis reveals a clear and physically grounded hierarchy: NIR and SWIR channels are the primary drivers of change detection and regression performance, consistent with their known sensitivity to water absorption and surface moisture. Conversely, RGB channels contribute more to stable-region prediction, likely because they capture land cover context that anchors no-change predictions. The notable inversion of NIR saliency between negative and positive change classes (where NIR aids NEG detection but affects POS detection) points to a non-trivial spectral interaction that needs further investigation with domain experts, possibly related to vegetation responses confounding water surface signals.

Climate variability and systematic failure modes. The climate subgroup discovery reveals that model performance is not uniformly distributed across environmental regimes: certain combinations of climate conditions consistently produce larger performance divergences, indicating that ACTU learns representations that generalize better to some hydrological contexts than others. This could be a consequence of the geographic and climatic composition of the training data, where some regimes are more represented than others, or of complex hydrological interactions that are not fully captured by the included input variables.

Implications of the composite regression loss. The ablation on the regression loss component shows that the combination of multiscale and wavelet losses achieves a better precision-recall trade-off than either component alone. The multiscale term promotes spatially coherent predictions at coarse resolution, reducing false positives from noise, while the wavelet term penalizes errors in high-frequency structures such as shoreline boundaries.

Broader applicability. While ACTU is evaluated on lakes and rivers across Europe, North America, and South America, the modular architecture allows the model to be deployed in other settings with only optical imagery. Additionally, the pretraining strategy on Landsat-5 and fine-tuning on Sentinel-2 demonstrates a viable transfer learning pathway across sensor generations, which is of practical relevance as new satellite missions come online.

9 Limitations

Despite the advances presented, several limitations should be acknowledged to guide future research.

Dataset Construction. While our heuristic selection provides a scalable way to filter all possible areas around the globe, it is important to mention that not every patch must contain a meaningful change: while the median change value is strongly skewed toward zero, it is relevant to note this reflects the real rarity of strong changes, while subtle or no change is more common. Since only lakes ≥ 100 km^2 and rivers with a Strahler flow order ≤ 4 are included, smaller water bodies that could be ecologically and agriculturally significant are excluded. Additionally, the dataset relies exclusively on optical imagery. While they are spectrally rich, cloud cover and atmospheric interference remain fundamental obstacles. Synthetic Aperture Radar (SAR) data (which is cloud-independent) could be included as additional information.

Geographic and climatic coverage. HydroChronos covers Europe, the United States, and Brazil, which, while geographically diverse, leaves significant hydrological regimes underrepresented (e.g., arid regions of Africa and Central Asia, monsoon-driven systems of South and Southeast Asia, and Arctic/subarctic permafrost areas). The concentration of lakes in the US and Europe and the exclusive focus on rivers in Brazil introduces a structural imbalance between the fine-tuning and test splits that may affect generalization to tropical lake systems. Future extensions of the dataset should prioritize coverage of these underrepresented regions.

Temporal resolution and sensor gap. The dataset relies on yearly composites from two distinct sensors (Landsat-5 and Sentinel-2), with a deliberate exclusion of the 2011–2014 period to avoid Landsat-7 striping artifacts. This gap interrupts temporal continuity and prevents the modeling of dynamics occurring in that interval. Additionally, restricting image acquisition to the May–August window (chosen to minimize seasonal variability in the Northern Hemisphere) introduces a systematic bias that may miss important flood or drought cycles that peak outside this window.

DEM Limitations. The Copernicus GLO-30 DEM is a static, single-timestep layer derived from TanDEM-X acquisitions collected between 2011 and 2015 [3]. Terrain changes occurring outside this window, such as reservoir construction, large-scale land subsidence, or sedimentation, are not captured.

MNDWI-based target. The use of MNDWI as an annotation-free proxy for water dynamics is pragmatic at scale but introduces well-known confounds: ice, snow, turbid water, and certain vegetation types can produce MNDWI signals that mimic or suppress water extent changes. While the median-based target aggregation reduces single-image noise, systematic confounds at the patch level (e.g., perennial ice cover at high-latitude lakes) cannot be fully eliminated. The fixed global threshold $t = 0.1$ further means that the definition of "relevant change" is uniform across climatically and ecologically heterogeneous settings. Adaptive or region-specific thresholds may better reflect real-world significance but require expert validation.

Climate data approximation. TERRACLIMATE provides climate data at approximately 4.6 km spatial resolution, significantly coarser than the 30 m satellite imagery. Beyond resolution, several variables are themselves outputs of a simplified linear water balance model rather than direct observations, which may introduce systematic errors. Higher-resolution reanalysis products (e.g., ERA5-Land) or in situ gauge data could provide more physically consistent climate forcing in future work.

XAI analysis. The XAI analysis also leaves room for further refinement. Local subgroup explanations require inspecting one subgroup at a time, providing a detailed but inherently local view of model behavior; yet global Shapley analysis complements this perspective by aggregating contributions across all discovered subgroups. Moreover, subgroup discovery depends on the discretization of continuous climate variables. In this work, we use frequency-based bins to obtain balanced groups, but domain-expert thresholds could define more physically meaningful climate regimes and further improve interpretability. Finally, the per-channel saliency analysis evaluates each input channel independently. Future work could extend the XAI framework to account for feature interactions and joint modality effects.

Model architecture and scalability. ACTU employs ConvLSTM for temporal aggregation, which is more suited and efficient for short sequences, but it may not scale well to longer historical contexts or higher-frequency temporal inputs. Attention-based temporal encoders could offer better scalability if there is any need for longer time series.

Additionally, the current evaluation considers only a single-step forecast ($T+1$ period). Multi-step or probabilistic forecasting, which is more relevant for operational water resource management, remains an open direction.

Hyperparameter sensitivity. The composite regression loss L_T introduces additional hyperparameters (λ, α, wavelet weights W) that were set based on preliminary experiments and domain priors. As noted in the ablation study, a more exhaustive search could yield further improvements. Similarly, the climate variable subset of five was chosen based on domain knowledge and multicollinearity reduction, but a systematic feature selection procedure could validate or revise this choice.

Task Complexity. Regressing the signed change T (rather than |T|) proved unsuccessful in preliminary experiments and was deferred to future work, limiting the model's ability to produce signed change maps in a regression setting, which would be more directly interpretable for water gain/loss analysis in operational contexts.

10 Conclusion

Our findings lay a foundation for advancing predictive spatio-temporal modeling in hydrology and for fostering research in water resource management in an era of significant environmental change. HydroChronos provides the first large-scale, multi-modal dataset explicitly designed for long-term surface water dynamics forecasting, integrating decades of Landsat and Sentinel-2 imagery with climate variables and DEM data. ACTU establishes a multimodal baseline that combines visual features, climate variables, and elevation, and our detailed analysis of its architecture, training strategy, loss functions, and gating mechanism clarifies key design choices for spatio-temporal predictive modeling. Our comprehensive XAI study identifies salient spectral bands and quantifies the impact of climate variables, revealing both strengths and limitations of the model and providing actionable insights for future work.

In future work, we plan to investigate additional tasks (e.g., MNDWI regression), and develop models specifically tailored to extreme events. Integrating data from radar altimetry, such as SWOT, could provide water surface elevation, while in situ measurements from river gauges could complement spectral data and improve physical consistency.

Acknowledgments

This study was partially supported by the "WEBFARE" (Nr. FISA2022-00908), funded by the Italian Ministry of University and Research under the FISA 2022 programme (D.D. No. 1405 of 13/09/2022,

Fondo Italiano per le Scienze Applicate – FISA). This manuscript reflects only the authors' views and opinions; the Ministry cannot be held responsible for them.

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] Ingrid Daubechies. 1992. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics. doi:10.1137/1.9781611970104

[13] 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

[14] 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

[15] 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

[16] Ayan S Fleischmann, Fabrice Papa, Stephen K Hamilton, Alice Fassoni-Andrade, Sly Wongchuig, Jhan-Carlo Espinoza, Rodrigo C D Paiva, John M Melack, Etienne Fluet-Chouinard, Leandro Castello, Rafael M Almeida, Marie-Paule Bonnet, Luna G Alves, Daniel Moreira, Dai Yamazaki, Menaka Revel, and Walter Collischonn. 2023. Increased floodplain inundation in the Amazon since 1980. *Environmental Research Letters* 18, 3 (Feb. 2023), 034024. doi:10.1088/1748-9326/acb9a7

[17] 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

[18] 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

[19] 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

[20] 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

[21] 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

[22] 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

[23] 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

[24] 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

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

[26] S.G. Mallat. 1989. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11, 7 (1989), 674–693. doi:10.1109/34.192463

[27] 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

[28] 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

[29] 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

[30] Eliana Pastor, Luca De Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the 2021 International Conference on Management of Data*. 1400–1412.

[31] 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.

[32] 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

[33] 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.

[34] 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 *Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25)*. ACM, 265–276. doi:10.1145/3748636.3762732

[35] 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

[36] 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

[37] 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

[38] 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

[39] 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

[40] Lloyd S Shapley. 1953. A value for n-person games. *Contributions to the Theory of Games* 2, 28 (1953), 307–317.

[41] 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

[42] 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

[43] 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

[44] 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

[45] 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

[46] 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

[47] 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

[48] 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

[49] 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

[50] 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

[51] 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

[52] 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

[53] 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

[54] 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

[55] 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

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