Session NH1.3
Hydrologic Extremes: Spatiotemporal Evolution, Cascading Impacts, and Adaptive Resilience

EGU26-9856

# ENSO impacts on flood risk and insurance claims in the United States: a machine learning approach

[1*]Konstantinos-Christofer Tsolakidis, [1]Konstantinos Papoulakos, [1]Nikolaos Tepetidis, [1]Theano Iliopoulou, [1]Panayiotis Dimitriadis, [2]Dimosthenis Tsaknias and [1]Demetris Koutsoyiannis

DOI: https://doi.org/10.5194/egusphere-egu26-9856

[1]Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Heroon Polytechneiou 5, GR-157 80 Zografou, Greece
[2] Independent researcher, Greece
* Corresponding author. E-mail address: kctsola@gmail.com

# Highlights

This research investigates the influence of the **El Niño–Southern Oscillation (ENSO)** on extreme **flood** events in the United States and its potential connection to **actual flood insurance claims**.

Integration of multiple datasets, including ENSO indices from **NOAA**, **US-CAMELS** streamflow data, **FEMA NFIP** claims data, **COBE** sea surface temperature (SST), **digital elevation models** (DEM), **National Hydrography Dataset** (NHD), **OpenStreetMap** (OSM), and **US Census data.**

The results indicate that correlations between **ENSO indices and streamflow** data are stronger than those between ENSO indices and insurance claim records, which highlights the influence of **socioeconomic factors** on the insurance claim filing process.

A machine learning model was developed to **predict** flood insurance claims per 100,000 residents.

The study concludes that ENSO indices can contribute to flood risk prediction frameworks.

**Keywords**: Flood insurance claims, Streamflow extremes, Climate variability, Machine Learning, FEMA

# Abstract

This research investigates the influence of the El Niño–Southern Oscillation (ENSO) on extreme flood events in the United States and its potential connection to **flood insurance claims** from the FEMA National Flood Insurance Program (NFIP). Given the recently observed increase in the frequency of extreme weather events, this study aims to quantify the correlation between **ENSO indicators** and recorded **economic losses** at state and county levels across the USA. Emphasis is particularly placed on the state of California, which is highly sensitive to El Niño events.

The methodology is based on the integration of multiple datasets, including ENSO indices from NOAA, US-CAMELS streamflow data, COBE sea surface temperature (SST), digital elevation models (DEM), National Hydrography Dataset (NHD), OpenStreetMap (OSM), and US Census data. From these datasets, geospatial and physical features were extracted, such as hydrographic and road network density, mean elevation, distance to the coastline, county centroid coordinates, and population. These features were analyzed using statistical tools, including the Pearson correlation coefficient and Threshold Exceedance Analysis, applied across multiple percentile showing thresholds (90–99%).

In addition, a **machine learning model** was developed to predict flood insurance claims per 100,000 residents. The results indicate that correlations between ENSO indices and streamflow data are significantly stronger than those between ENSO indices and insurance claim records, highlighting the substantial influence of **socioeconomic factors** on the insurance claim filing process. California exhibits the highest positive correlation between the maximum annual ENSO index and insurance claims ($r \approx 0.35$). The developed **CatBoost model** can be used to predict a high percentage (>60%) of their variability, using both static and dynamic features.

The study concludes that ENSO indices can contribute meaningfully to flood risk prediction frameworks. Future work will focus on extending the analysis to additional states or the entire USA and incorporating new explanatory features to further improve model performance.

# Datasets

**Streamflow:**

- US-CAMELS (Newman et al., 2014)

**Sea Surface Temperature:**

- COBE Sea Surface Temperature (Hirahara et al., 2014)

**Climate indices:**

- ENSO (NOAA, 2017)

**Flood insurance claims:**

- FEMA publishes NFIP claims and policy data (FEMA, 2019)

**Geospatial:**

- State and County Boundaries (TIGER/Line) – U.S. Census Bureau (2023)
- National Hydrography Dataset (NHD) – U.S. Geological Survey (USGS, 2023)
- Road Network Data (OSM / NTD) – OpenStreetMap contributors (2023)
- Digital Elevation Model (DEM) – U.S. Geological Survey 3D Elevation Program (3DEP) (USGS, 2023)
- Population Estimates Program (PEP) – U.S. Census Bureau (2023)

# Methodology

The following techniques and methods were used:

- Moving Averages
- Pearson correlation coefficient
- Peak Over Threshold Analysis
- Mapping Analysis using QGIS
- Statistical Data Distribution
- Box Plot and Dumbbell Charts
- Collective risk model in flood insurance (Papoulakos et al., 2025)

# Methodology: Machine Learning Model Training

A regression-based machine learning approach was adopted, aiming to estimate the annual number of flood insurance claims in the counties of California.

The data were divided into two sets according to the following ratio, which represents a standard choice:

- Training set: 80% of the data (1978–2006, 2012–2024)
- Test set: 20% of the data (2007–2011)

The performance of the model was evaluated using three main statistical metrics:

- $R^2$, namely the coefficient of determination
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)

# Methodology: Exploration of Machine Learning Models & Performance Optimization

As part of the optimization process, modifications were tested both in terms of the input variables and the configuration of the applied algorithms, with the primary objective of improving model performance ($R^2$). The following models were evaluated:

- Random Forest (RF) ($R^2$ = 0.090).
- Random Forest with an expanded set of input variables. Additional important dynamic and static features were incorporated, including ENSO indices, population, mean elevation, as well as geographical characteristics such as County centroid coordinates and distance from the coastline ($R^2 \approx 0.26$).
- XGBoost (Chen and Guestrin, 2016), where the input variables were defined as static and dynamic datasets, and the hyperparameters were carefully tuned to maximize model performance for the specific problem under investigation. This resulted in an increase in predictive performance ($R^2 \approx 0.39$), highlighting the importance of hyperparameter tuning in complex machine learning problems to achieve optimal model efficiency (Tepetidis et al., 2024).
- CatBoost ($R^2$ = 0.638, RMSE = 0.81, and MAE = 0.59).

# Methodology: The Catboost Machine Learning Algorithm (I)

For our model, the CatBoost algorithm was selected, as it offers significant advantages, including:

- Automatic detection of interactions between features, while maintaining high accuracy.
- It outperformed other models of the same category, as it rarely exhibits prediction shift.
- Through the balanced correction applied across all trees, overfitting to the data is effectively avoided.

# Methodology: The Catboost Machine Learning Algorithm (II)

The model achieved an $R^2$ value of 0.638, indicating that it can explain more than 60% of the variance of the dependent variable. This value is considered satisfactory for the specific problem addressed in this study, given the high level of complexity and heterogeneity of the input variables. At the same time, the remaining evaluation metrics yielded values of 0.81 for RMSE and 0.59 for MAE.

Overall, considering all three evaluation metrics together, the results confirm that the model demonstrates balanced and reliable performance.

# Methodology: Machine Learning Model

Having compiled a comprehensive list of all variables for each year across all California Counties, alongside the ENSO index, we ultimately obtained a database of 1,572 entries with 8 differential variables (both static and dynamic).

Static features:

- Hydrographic Network Density
- Road Network Density
- County Centroid Coordinates
- Centroid Distance from Coastline Mean Elevation (or Average Altitude)
- Population (2023)

Dynamic features:

- ENSO Index (El Niño-Southern Oscillation Index)
- Number of Insurance Claims

# Results:
# Correlations Between ENSO, Insurance Claims and occurrence of extreme flood events

# Correlations between ENSO and actual flood insurance claims in the USA

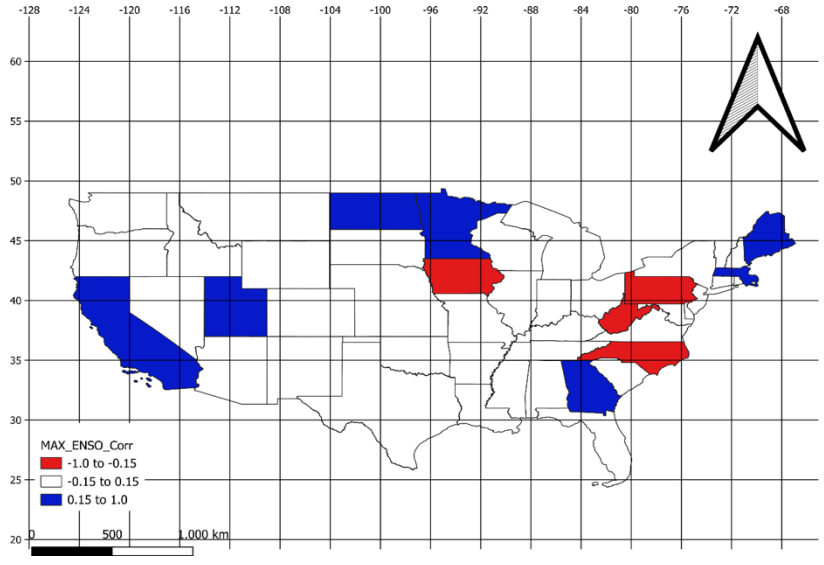


**Fig. 1** Spatial correlation of maximum annual ENSO index and flood insurance claims in the United States (1980–2024)

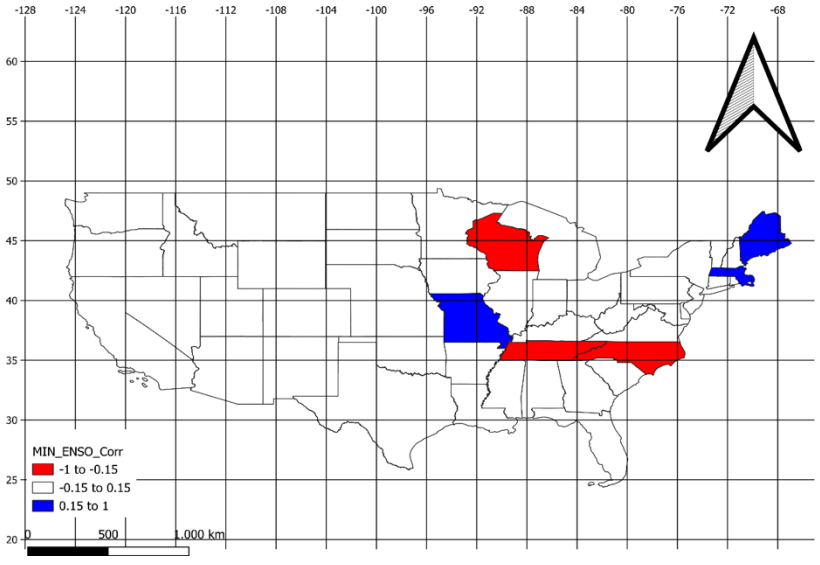


**Fig. 2** Spatial correlation of minimum annual ENSO index and flood insurance claims in the United States (1980–2024)

Significant spatial variability is observed, reflecting the non-uniform influence of ENSO on hydrological risk across different regions (Kunkel et al., 2003; Hamlet & Lettenmaier, 2007).

In contrast, local characteristics such as terrain morphology and urbanization make the relationship between ENSO phenomena and flood-related damages more complex (Hamlet & Lettenmaier, 2007).

# Correlations between ENSO and actual flood insurance claims in the state of California (I)

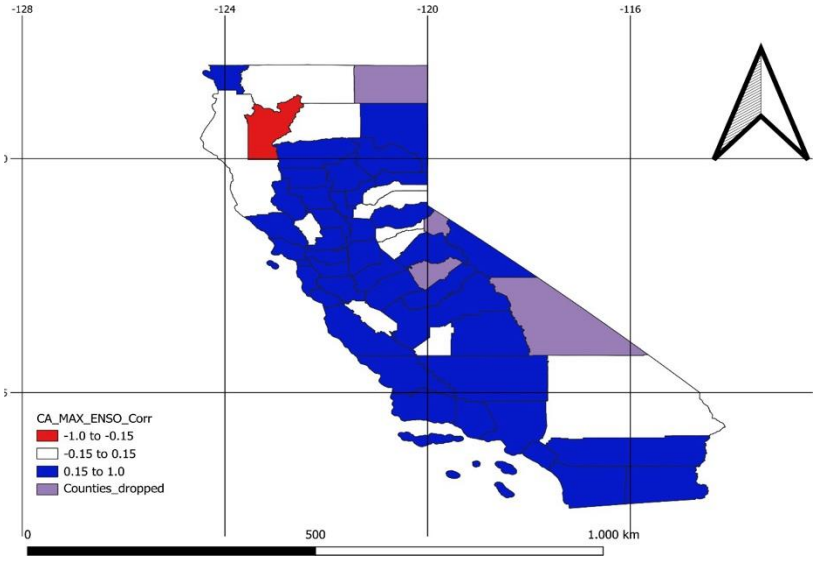


**Fig. 3** Spatial correlation of maximum annual ENSO index and flood insurance claims in the state of California (1980–2024)

It is observed that most counties exceed the upper threshold of +0.15 (blue shading), while all remaining counties fall within the intermediate category (-0.15 to 0.15). An exception is Trinity County, where the primary issue is snowfall rather than ENSO-related impacts, as confirmed in the relevant literature (U.S. Geological Survey, 1998).

This finding reinforces the argument previously presented, namely that the number of insurance claims is not solely (although largely) influenced by ENSO phenomena, but also by a range of other factors such as topographic and climatic conditions, the geographical location of each region, and other local characteristics.

# Correlations between ENSO and actual flood insurance claims in the state of California (II)

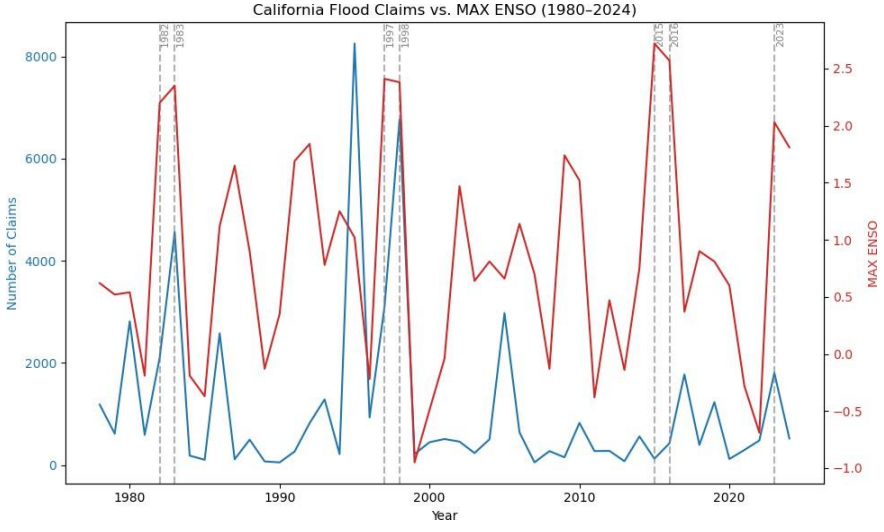


**Fig. 4** Correlation between insurance claims and maximum ENSO index values in California

The correlation supports the hypothesis that climate variability through ENSO influences the frequency and severity of flood events. Although a corresponding increase in insurance claims is not always observed for each ENSO peak—likely due to local factors such as infrastructure, precipitation patterns, and soil characteristics—the overall pattern suggests a significant influence.

Additionally, it is worth referencing the study by Murakami et al. (2025), which concludes that ENSO phenomena are not, in themselves, a sufficient indicator of risk, as the presence of dry air masses (e.g., Saharan dust), atmospheric stability, and other regional factors can suppress the development and intensity of such events.

# Correlations between between ENSO phenomena and the occurrence of extreme flood events (I)

- The potential relationship between ENSO phenomena and the occurrence of extreme flood events is examined.
- To this end, discharge time series from the US-CAMELS dataset (Newman et al., 2014) were analysed using the peak-over-threshold method.
- More specifically, for each of the considered gauging stations, the collective risk *S* (Papoulakos, 2025) was derived, along with the number of threshold exceedances.
- In greater detail, Pearson correlation coefficients were calculated between two key flood risk variables derived from the discharge time series—the number of threshold exceedances (*N*) and the collective risk—and three main indices used to quantify ENSO phenomena: the annual mean, annual maximum, and annual minimum values.
- The analysis was performed for four successive discharge percentile thresholds (90%, 95%, 98%, and 99%), corresponding to increasing levels of event intensity observed at the studied stations.

# Correlations between between ENSO phenomena and the occurrence of extreme flood events (II)

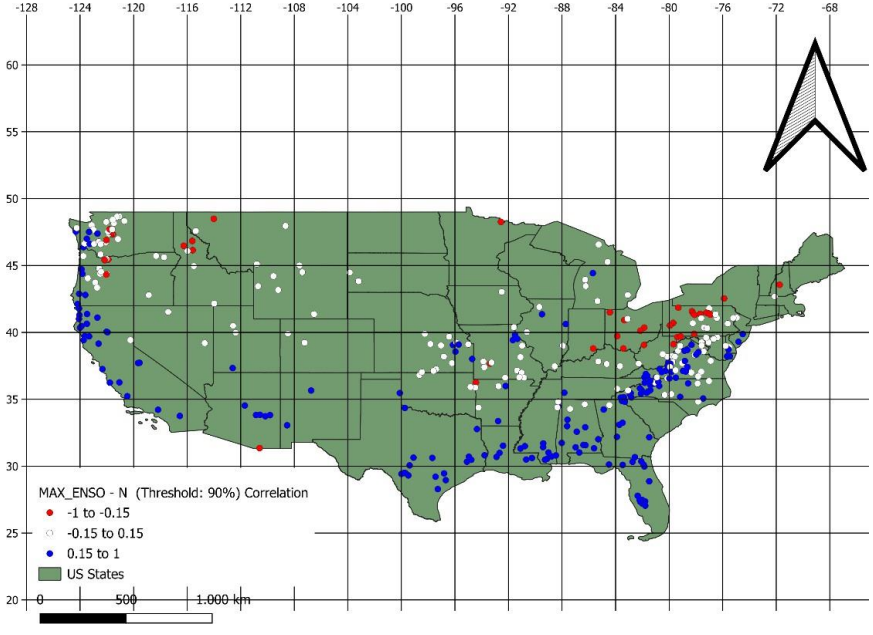


**Fig. 5** Map of the correlation degree of exceedances (N_SUM) for the 90% exceedance threshold compared with the maximum ENSO index values (MAX_ENSO) by US-CAMEL gauge location

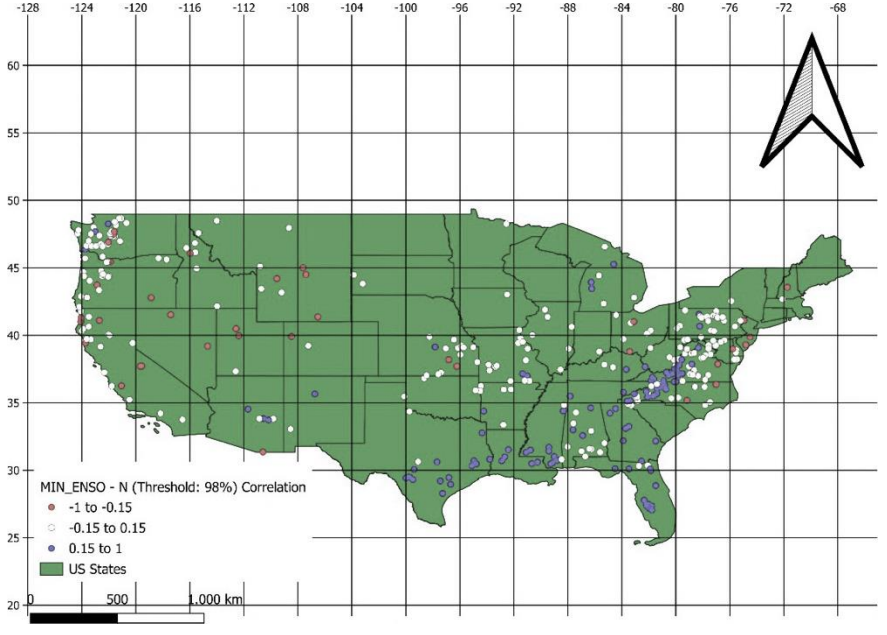


**Fig. 6** Map of the correlation degree of exceedances (N_SUM) for the 98% exceedance threshold compared with the maximum ENSO index values (MAX_ENSO) by US-CAMEL gauge location

# Correlations between between ENSO phenomena and the occurrence of extreme flood events (III)

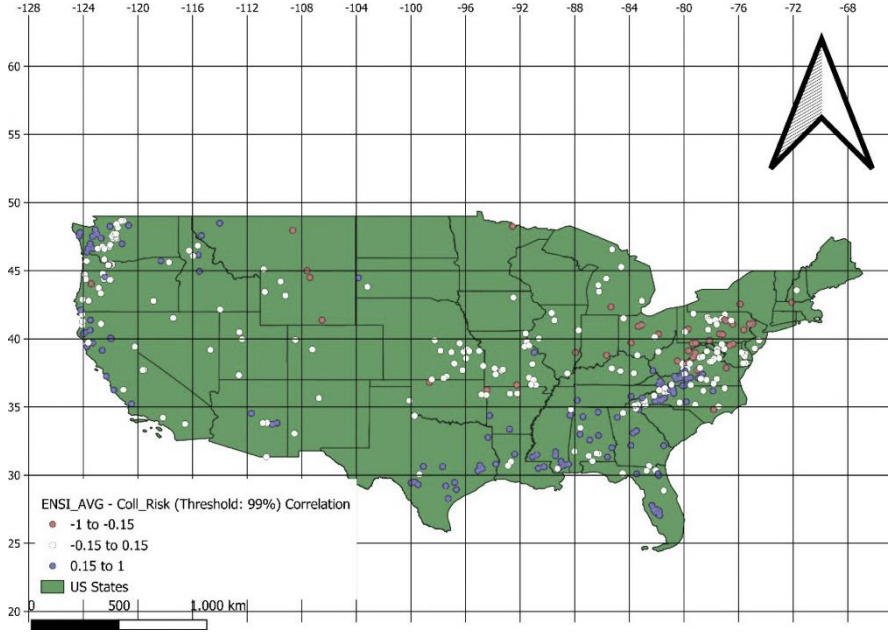


**Fig. 7** Map of the correlation degree of exceedances (N_SUM) for the 99% exceedance threshold compared with the maximum ENSO index values (MAX_ENSO) by US-CAMEL gauge location

By observing all three maps, it becomes evident that most coastal regions exhibit stronger correlations with ENSO indices. This is fully justified and consistent with the existing literature.

In the first map (Fig. 5), which refers to the correlation with the maximum annual ENSO index values (MAX_ENSO) and is associated with El Niño events, a particularly clear pattern emerges, with the vast majority of coastal stations showing correlation values greater than 0.15. Specifically, 171 out of the 360 stations exhibit a correlation coefficient above 0.15, a finding that supports the aforementioned hypothesis.

# Correlations between between ENSO phenomena and the occurrence of extreme flood events (IV)

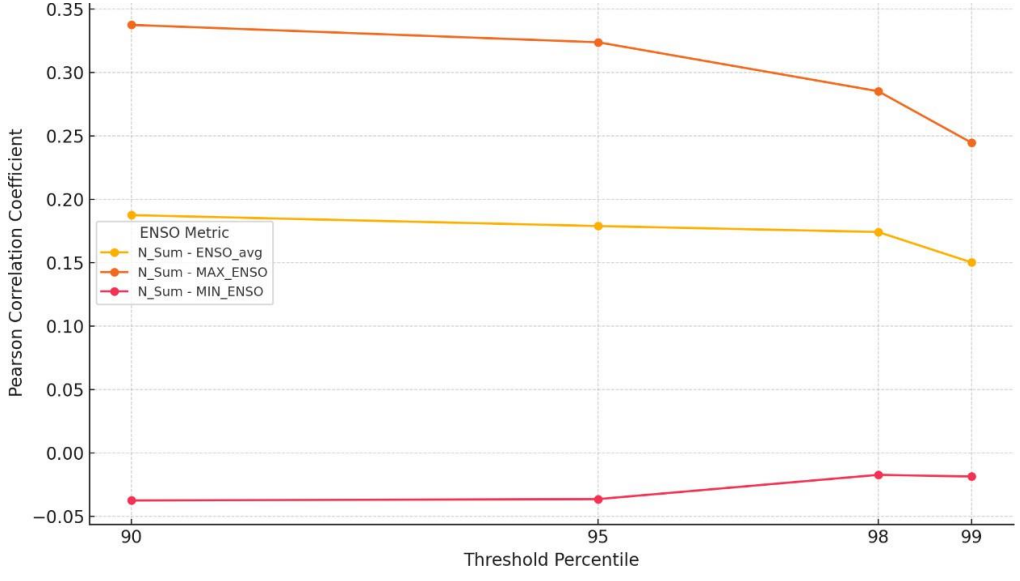


**Fig. 8** Variation of the Pearson correlation coefficient between flood risk variables (number of events) and ENSO indices (mean, maximum, minimum) across different extreme value thresholds (90%, 95%, 98%, 99%)

It is observed that as the threshold increases (i.e., when focusing on more extreme flood events), the correlation between the number of threshold exceedances and all ENSO indices decreases. Any correlation with the annual minimum ENSO value remains negative or negligible across all exceedance thresholds.

These results are consistent with previous studies that document a significant influence of ENSO phenomena on flood-related damages (Ward et al., 2014). The use of percentile-based thresholds for identifying extreme hydrological events is also well supported in the literature on extreme value analysis (Coles, 2001).

# Comparison of Correlations Between ENSO Indices, Insurance Claims, and Streamflow Threshold Exceedance Indicators (I)

Within the framework of the present study, we compared the correlations between three ENSO indices (MIN_ENSO, MAX_ENSO, ENSO_avg) and two different variables: streamflow indicators (peak-over-threshold metrics derived from the 360 gauge locations of the US-CAMELS dataset) and flood insurance claims (FEMA NFIP dataset).

The correlations with streamflow data generally exhibit higher values than the corresponding correlations with insurance claims, particularly for the MAX_ENSO index, indicating that the influence of ENSO phenomena is more readily reflected in hydrological records.

The correlation values derived from insurance claims are comparatively lower, which may be attributed to external socioeconomic factors or variations in local insurance policies. Overall, the results suggest that the impact of ENSO is more easily detectable in hydrological data than in recorded economic/insurance loss data.

# Comparison of Correlations Between ENSO Indices, Insurance Claims, and Streamflow Threshold Exceedance Indicators (II)

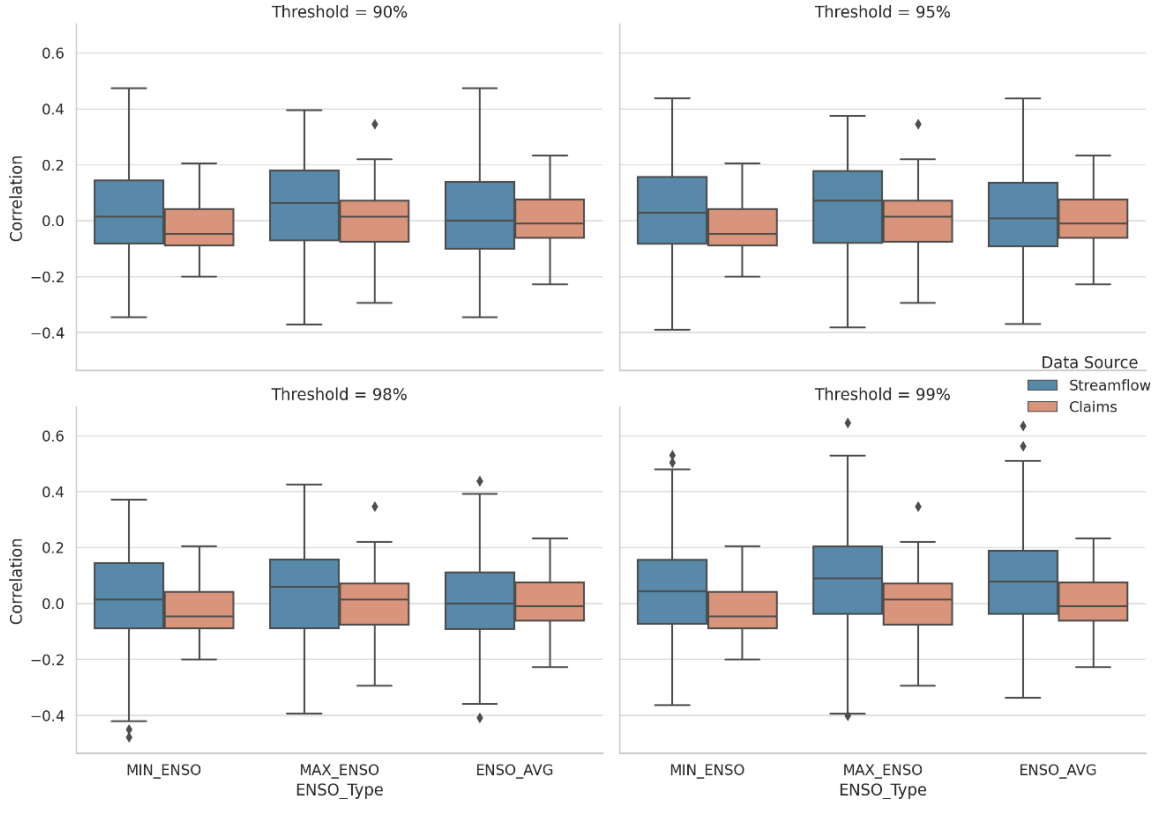


**Fig. 9** Comparison of Boxplot Diagrams of Correlations Between ENSO Indices (MComparison of Boxplot Diagrams of Correlations Between ENSO Indices (Mean, Maximum, Minimum), Insurance Claims, and Streamflow Threshold Exceedance Counts for Different Thresholds (90, 95, 98, 99%).

# Comparison of Correlations Between ENSO Indices, Insurance Claims, and Streamflow Threshold Exceedance Indicators (III)

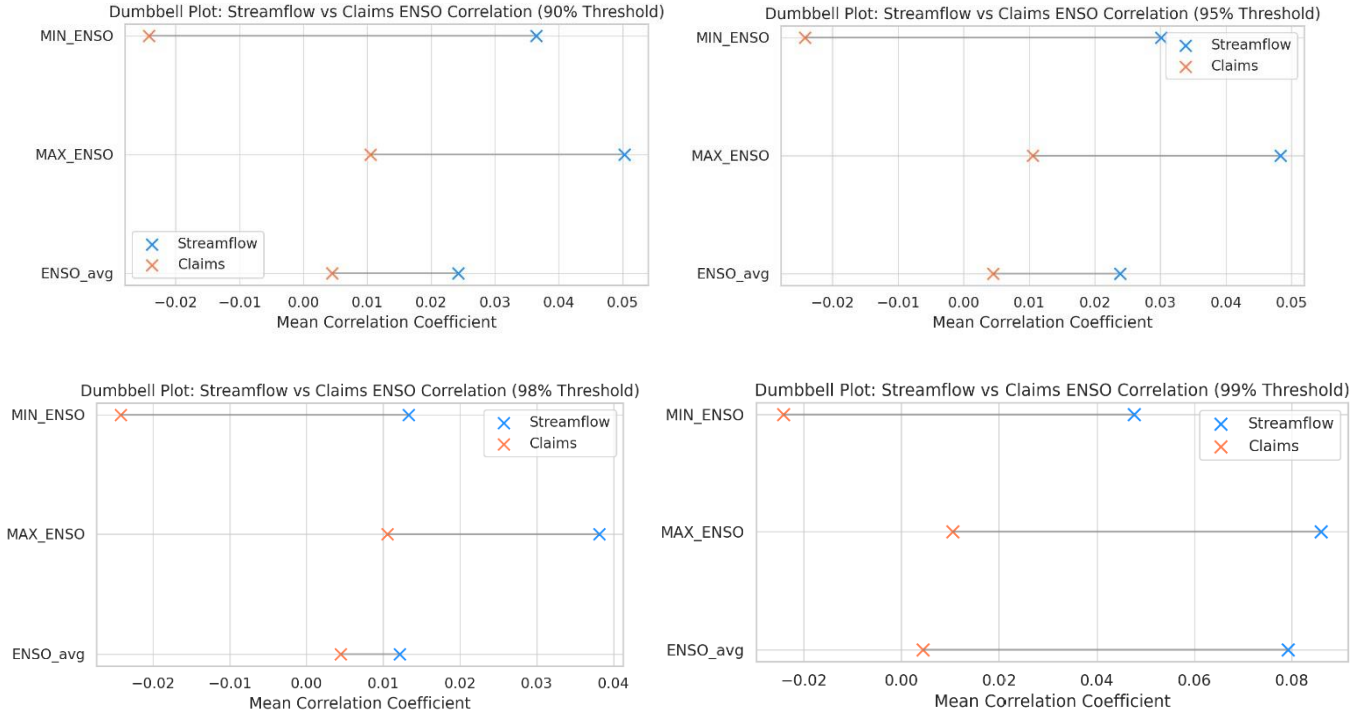


**Fig. 10** Comparison of Dumbbell Plots of Correlations Between ENSO Indices (Mean, Maximum, Minimum), Insurance Claims, and Streamflow Threshold Exceedance Counts for Different Thresholds (90, 95, 98, 99%)

# Machine learning model (I)

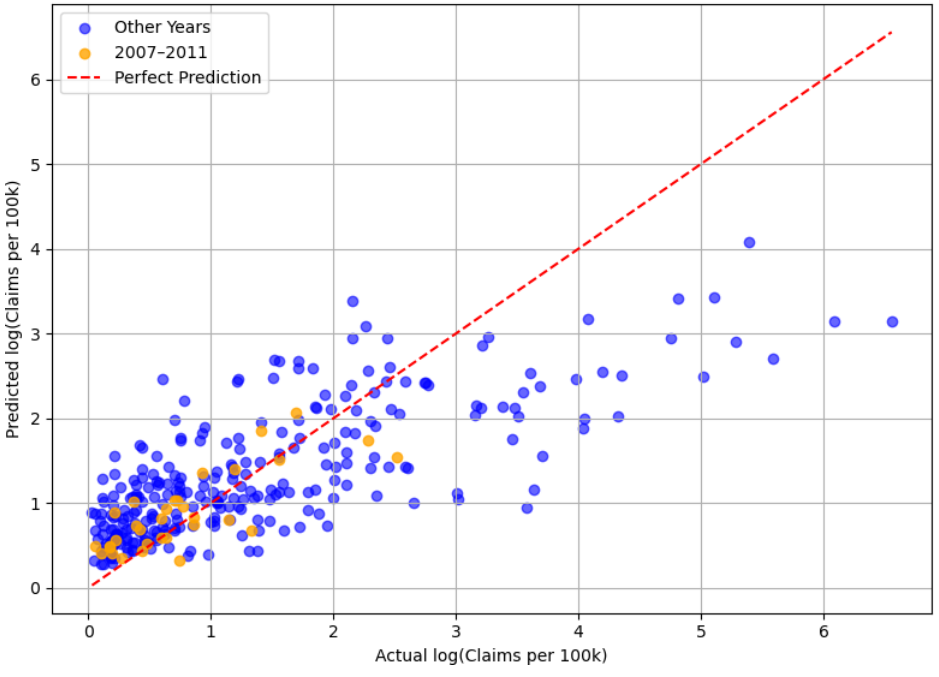


**Fig. 11** Scatter plot between predicted and observed values on a logarithmic (log) scale of claims per 100,000 residents for California

The final machine learning model was developed based on the CatBoost regression framework. To evaluate its performance, a scatter plot was created to compare predicted and observed values, using the logarithmically transformed number of insurance claims per 100,000 inhabitants as the target variable.

The red dashed line represents the 1:1 line, which corresponds to a perfect model and is used as a reference for comparing the model's predictions with the observed values. Predictions for the years 2007–2011 are highlighted in a different color, representing the five-year period selected as the target for presentation purposes within the context of this research. The model can, in principle, be applied to any desired target time period.

The model predictions show a positive trend and a reasonably good agreement with the observed values, although a slight underestimation is observed in the higher value range. The resulting model achieves a coefficient of determination of $R^2$ = 0.638, which is considered reasonably satisfactory.

# Machine learning model (II)

Our database consists of 1,572 records. In this study, in order to train the developed model, it was modified to include 1,418 initial observations, comprising all insurance claims recorded within the boundaries of the State of California from 1978 to 2024 (with the exception of the target five-year period 2007–2011).

In the final model formulation, the feature importance ranking was also evaluated using the PredictionValuesChange metric. More specifically, these values indicate the degree of influence of each variable on the model's predictive performance. The higher the value, the greater the impact of the corresponding feature on the model's predictions.

This evaluation is considered important for understanding the most influential factors contributing to insurance claims.

| | Feature | Importance |
|---|---|---|
| 1 | ENSO_index | 50.44 |
| 2 | Population | 20.69 |
| 3 | Centroid Longitude | 9.02 |
| 4 | Centroid Latitude | 8.05 |
| 5 | Distance_to_sea | 4.48 |
| 6 | Hydro_density | 3.11 |
| 7 | Elevation_mean | 2.15 |
| 8 | Transport_density | 2.06 |

**Table 1** Ranking Table of Feature Importance

| Metric | Value |
|---|---|
| $R^2$ | 0,64 |
| RMSE | 0,81 |
| MAE | 0,59 |

**Table 2** Performance Metrics of the Final Machine Learning Model

# Conclusions (I)

- Analysis of the research results showed that the correlations between ENSO indices and flood insurance claims were low.
- Specifically, 66% of the United States exhibited a correlation coefficient below 0.10 with the annual mean ENSO index (ENSO_AVG), corresponding to 33 states.
- Similarly, 56% of the states (28 states) showed a correlation coefficient below 0.10 with the annual maximum ENSO index (MAX_ENSO).
- In both cases, the remaining states presented correlation values higher than 0.10.
- These low correlations can be attributed to several factors:
    - ENSO indices are global-scale climate indicators and are not specifically designed to quantify impacts at the regional scale of the United States.
    - Flood insurance claims are not determined solely by meteorological conditions, but are also influenced by social and institutional factors, such as:
        - degree of urbanization,
        - insurance practices,
        - public awareness and preparedness, among others.
    - The influence of ENSO phenomena does not always result in floods and damages, but mainly during their peak intensity phases.

# Conclusions (II)

- Extensive comparisons were conducted between the correlations of ENSO indices with both the recorded total insurance claims and the corresponding streamflow variables.
- ENSO indices exhibit stronger correlations with streamflow data.
- This difference can be attributed to several factors:
    - Streamflow data primarily represent the physical hydrological phenomenon, where the influence of a climatological index may be more directly identifiable.
    - In contrast, insurance claims are strongly influenced by socioeconomic factors, such as the degree of urbanization, the population education level, the standard of living, and the existence of insurance culture, etc.
- A characteristic example is California, which represents a highly developed society where local insurance policies and citizens' insurance culture lead to increased numbers of insurance claims.
- On the other hand, in other states, ENSO-related impacts may be clearly evident, yet no insurance claims may ever be submitted. As a result, these figures may not accurately reflect the actual impacts, leading to lower correlations with ENSO indices when compared to the number of insurance claims.
- In conclusion, hydrological data provide a significantly clearer representation of the impacts and consequences of ENSO phenomena.

# References


Chen, T., & Guestrin, C. (2016) XGBoost: A scalable tree boosting system.In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). https://doi.org/10.1145/2939672.2939785

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer. https://doi.org/10.1007/978-1-4471-3675-0

FEMA (2019). FEMA publishes NFIP claims and policy data. https://www.fema.gov/openfema-data-page/fima-nfip-redacted-claims-v2

Hamlet, A.F., & Lettenmaier, D.P. (2007). Effects of 20th century warming and climate variability on flood risk in the western US. Water Resources Research, 43(6). https://doi.org/10.1029/2006WR005099

Hirahara, S., Ishii, M., & Fukuda, Y. (2014).Centennial-scale sea surface temperature analysis and its uncertainty.Journal of Climate, 27(1), 57–75. https://doi.org/10.1175/JCLI-D-12-00837.1

Kunkel, K.E., Andsager, K., & Easterling, D.R. (2003). Long-term trends in extreme precipitation events over the conterminous United States and Canada. Journal of Climate, 16(13), 2125–2147. https://doi.org/10.1175/2768.1

Murakami, H., Nakano, M., Doi, T., Ogata, T., Nasuno, T., Yamada, Y., Mizuta, R., Yoshida, K., and Zhang, W. (2025). Forecasting Challenges in 2024: Systematic Overprediction of North Atlantic Tropical Cyclone Activity in Seasonal Forecasts. Geophysical Research Letters (submitted).

Newman A, Sampson K, Clark MP, Bock A, Viger RJ, & Blodgett D (2014). A large-sample watershed-scale hydrometeorological dataset for the contiguous USA. UCAR/NCAR, Boulder, CO. https://doi.org/10.5065/D6MW2F4D

NOAA. El Niño / Southern Oscillation (ENSO), Southern Oscillation Index (SOI), National Centers for Environmental Information (NCEI). https://www.ncei.noaa.gov/access/monitoring/enso/

OpenStreetMap contributors. (2023). Planet dump retrieved from https://planet.openstreetmap.org.https://www.openstreetmap.org

Papoulakos, K., Iliopoulou, T., Dimitriadis, P., Tsaknias, D., & Koutsoyiannis, D (2025). Spatiotemporal clustering of streamflow extremes and relevance to flood insurance claims: a stochastic investigation for the contiguous USA. Nat Hazards 121, 447–484. https://doi.org/10.1007/s11069-024-06766-z

# References


Tepetidis, N., Koutsoyiannis, D., Iliopoulou, T., & Dimitriadis, P. (2024). Investigating the performance of the Informer model for streamflow forecasting. Water, 16(20), Article 20734441. https://doi.org/10.3390/w16202041

U.S. Census Bureau. (2023). Population Estimates Program (PEP). United States Department of Commerce. https://www.census.gov/programs-surveys/popest.html

U.S. Census Bureau. (2023). TIGER/Line Shapefiles. United States Census Bureau.

U.S. Geological Survey (USGS). (2023). 3D Elevation Program (3DEP) – 1/3 arc-second historical elevation data. U.S. Department of the Interior. https://www.usgs.gov/3d-elevation-program https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

U.S. Geological Survey (USGS). (2023). National Transportation Dataset (NTD). U.S. Department of the Interior. https://www.usgs.gov/national-transportation-dataset

U.S. Geological Survey (USGS) (2023). National Hydrography Dataset (NHD). U.S. Department of the Interior. https://www.usgs.gov/national-hydrography

Ward, P.J., Jongman, B., Kummu, M., Dettinger, M.D., Weiland, F.S., and Winsemius H.C. (2014). Strong influence of El Niño Southern Oscillation on flood risk around the world. Proceedings of the National Academy of Sciences, 111(44), 15659–15664. https://doi.org/10.1073/pnas.1409822111