A Hybrid GPR-GAM Model for Enhanced Spatio-Temporal Climate Prediction in Kenya

Abstract

Climate change presents growing challenges in regions like Kenya, where diverse terrain and climatic variability complicate accurate environmental forecasting. Traditional climate models often fall short in capturing both the non-linear relationships among climatic variables and the spatial dependencies inherent in such data. To address these limitations, this study introduces a novel hybrid model that integrates Gaussian Process Regression (GPR) and Generalized Additive Models (GAM) to enhance spatio-temporal climate prediction. The model was developed by combining the structured, interpretable components of GAM with the spatially aware, probabilistic strengths of GPR, using climate data collected from the Google Earth Engine covering the period 2015–2024. Model parameters were estimated through generalized cross-validation and optimized using the L-BFGS algorithm. Results indicate that the hybrid model significantly improves predictive accuracy compared to standalone GPR or GAM approaches, achieving an RMSE of 1.27℃ and an R^2 of 0.91. These findings demonstrate the model's effectiveness in capturing Kenya's spatial and climatic heterogeneity. The study recommends the hybrid model's application in climate-sensitive sectors such as agriculture, infrastructure development, and early warning systems, with future work focusing on scalability and real-time deployment.

Keywords: Hybrid GPR-GAM; spatio-temporal climate prediction

1 Introduction

Climate variability and climate change present some of the most pressing challenges facing developing countries in the 21st century (Thornton et al. (2014)). Nowhere is this more apparent than in Kenya, a country marked by pronounced climatic heterogeneity and ecological diversity. From the arid and semi-arid regions of the north to the humid coastal lowlands and the temperate central highlands, Kenya's landscape presents complex environmental gradients that give rise to spatially and temporally dynamic weather patterns. These climatic variations have far-reaching implications for livelihoods, agriculture, public health, and national development goals. Agriculture, which accounts for a significant portion of Kenya's GDP and employs a majority of the population, is particularly vulnerable to temperature shifts and rainfall variability (Herrero et al. (2010)). Additionally, climate-sensitive sectors such as water resources, food security, infrastructure, and disaster risk management require precise and location-specific climate predictions to make informed decisions (Lawrence et al. (2023)). In this context, climate modeling emerges not merely as an academic exercise but as a crucial tool for resilience planning and adaptive policy design.

Conventional modeling approaches—such as General Circulation Models (GCMs) and Regional Climate Models (RCMs)—have proven valuable in assessing long-term climate trends and in providing boundary conditions for understanding global climate dynamics (Chokkavarapu and Mandla (2019)). However, these models typically operate at spatial resolutions too coarse to capture local-scale phenomena relevant to community-level planning in Kenya. Their reliance on rigid parametric assumptions and their limited ability to reflect microclimatic influences such as orographic effects, land-use change, and spatial heterogeneity further limit their applicability (Pathirana et al. (2014)). Downscaled versions of these models improve spatial resolution but still often fail to account adequately for the nonlinear, spatially structured relationships that define climate behavior in topographically diverse settings like Kenya.

Emerging statistical and machine learning models have been adopted to overcome some of these shortcomings, offering greater flexibility and adaptability to regional contexts (Wilson and Anwar (2024)). Among the most notable approaches are Generalized Additive Models (GAMs) and Gaussian Process Regression (GPR). GAMs are an extension of traditional linear models that allow for smooth, non-linear relationships between covariates and responses (Wiley et al. (2019)). They are especially useful for modeling variables like temperature that respond non-linearly to environmental gradients such as elevation and precipitation. GAMs also offer the advantage of interpretability: each term in the model can be examined independently to understand how specific factors contribute to the outcome (Zschech et al. (2022)). However, despite their transparency and computational efficiency, GAMs do not inherently account for spatial dependencies in the data, nor do they quantify the uncertainty associated with predictions—two critical features in the context of climate modeling and risk assessment.

In contrast, Gaussian Process Regression (GPR) is a non-parametric, Bayesian approach that excels in modeling spatial and spatio-temporal data (Pipia et al. (2021)). GPR offers probabilistic predictions, meaning it not only provides point estimates but also quantifies the uncertainty associated with those estimates. This is particularly important in regions like northern Kenya, where data sparsity can undermine confidence in deterministic models. GPR models incorporate spatial structure directly through kernel functions—such as the Matérn kernel—which can flexibly adapt to varying spatial scales and smoothness levels (Tuia et al. (2018)). However, GPR comes with its own set of challenges. It is computationally expensive, particularly for large datasets, and the influence of individual covariates is embedded within the kernel function, making the model less transparent and interpretable.

The limitations of using GAM or GPR in isolation highlight the need for a hybrid modeling framework that capitalizes on the strengths of both. GAM can capture structured, non-linear relationships between climatic variables and temperature, while GPR can model the residual spatial autocorrelation not explained by those variables. Together, these models can form a more comprehensive system capable of handling the structural and stochastic complexity inherent in climate data. Such an integrated approach is especially well-suited for countries like Kenya, where diverse terrain and climate interactions demand both interpretability and precision in predictive models.

This study introduces and evaluates a hybrid modeling framework that integrates GAM and GPR for spatio-temporal temperature prediction across Kenya. The hybrid model leverages the ability of GAM to model structured nonlinear effects such as elevation, precipitation, and temporal trends, and combines it with the strength of GPR to capture spatial dependencies and provide predictive uncertainty. The hybrid framework first uses GAM to estimate the systematic effects of covariates and then applies GPR to model the residuals, thereby incorporating spatial structure into the final prediction.

2 Materials and Methods

This study proposes a hybrid climate modeling framework that integrates Generalized Additive Models (GAM) and Gaussian Process Regression (GPR) to predict average daily temperature across Kenya using spatio-temporal climate data. The methodological workflow comprises four key components: data acquisition and preprocessing, model specification, parameter estimation and optimization, and performance evaluation.

2.1 Data Sources and Preprocessing

Daily climate and topographic data from 2015 to 2024 were obtained from Google Earth Engine (GEE). Temperature and precipitation variables were extracted from the ERA5-Land dataset, while elevation data were obtained from the Shuttle Radar Topography Mission (SRTM). A set of 500 spatial locations was randomly sampled across Kenya, ensuring representative coverage across diverse ecological zones such as highlands, coastal areas, and arid regions.

The dataset included the following variables: average daily temperature (T), precipitation (P), elevation (E), latitude (x_1), longitude (x_2), and time (t). Preprocessing steps involved outlier removal based on interquartile range thresholds, normalization of continuous covariates, and conversion of dates into numeric time indices. The data were split into training (80%) and test (20%) sets, ensuring spatial balance in the split.

2.2 Generalized Additive Model (GAM)

The first component of the hybrid model is a GAM that captures smooth, nonlinear relationships between temperature and selected covariates (elevation, precipitation, and time). The GAM is expressed as:

$$\eta(y) = \beta_0 + f_1(t) + f_2(E) + f_3(P) \tag{2.1}$$

where:

- β_0 is the intercept,
- $f_1(t)$ is a smooth function of time,

- $f_2(E)$ is a smooth function of elevation,
- $f_3(P)$ is a smooth function of precipitation.

The functions f_i are estimated using penalized thin plate regression splines. The residuals from the GAM model, $r(x)$, are computed as:

$$r(x) = T(x) - \eta(y) \tag{2.2}$$

These residuals are then modeled using GPR to account for spatial autocorrelation not captured by the GAM.

2.3 Gaussian Process Regression (GPR)

The GPR component models the residual spatial structure of temperature data. Let $r(x)$ denote the residual at location $x = (x_1, x_2)$ in spatial coordinates. GPR assumes that:

$$r(x) \sim \mathcal{GP}(0, K(x, x')) \tag{2.3}$$

where $K(x, x')$ is the covariance function (kernel) between locations x and x'. The Matérn kernel was selected for its flexibility and is defined as:

$$K_\nu(d) = \sigma^2 \cdot \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu} d}{\ell} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu} d}{\ell} \right) \tag{2.4}$$

where:

- $d = \|x - x'\|$ is the Euclidean distance between spatial points,
- ν is the smoothness parameter (set to 1.5),
- ℓ is the length scale,
- σ^2 is the signal variance,
- K_ν is the modified Bessel function of the second kind.

The GPR predictions are derived by maximizing the log-marginal likelihood:

$$\log p(\mathbf{r}|X) = -\frac{1}{2} \mathbf{r}^\top (K + \sigma_n^2 I)^{-1} \mathbf{r} - \frac{1}{2} \log |K + \sigma_n^2 I| - \frac{n}{2} \log 2\pi \tag{2.5}$$

where σ_n^2 is the noise variance and K is the covariance matrix over all spatial locations.

2.4 Hybrid GPR-GAM Framework

The final model is a summation of the structured GAM component and the spatial GPR component, along with an i.i.d. noise term:

$$T(x) = \eta(y) + f_{\text{GPR}}(x) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2.6}$$

Parameter estimation was performed in two stages: (1) fitting the GAM component using penalized least squares and selecting the smoothing parameter λ via generalized cross-validation (GCV), and (2) training the GPR component on the residuals using the L-BFGS optimization algorithm for hyperparameter tuning.

2.5 Model Evaluation

The performance of the hybrid model was evaluated against standalone GAM and GPR models using a held-out test set. Evaluation metrics included root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and coverage probability of 95% prediction intervals. Spatial diagnostics such as Moran's I and semivariograms were used to assess residual spatial dependence.

All computations were implemented in Python using packages such as `pyGAM`, `GPy`, `scikit-learn`, and `geopandas`. Visualization was conducted using `matplotlib` and `seaborn`.

3 Results and Discussion

The temperature data was collected from GEE for a period of 10 years (2015-2024). In order to achieve parsimonious results, summary tables were created, graphs were drawn, and the results were extensively discussed. The analysis was performed using R and Python software. The hybrid GAM-GPR model was designed to leverage the deterministic interpretability of GAM with the probabilistic spatial adaptability of GPR, in order to model the average temperature in Kenya with both structural precision and spatial stochastic flexibility.

3.1 Model's Predictive Performance

The implementation of the hybrid Generalized Additive Model–Gaussian Process Regression (GAM–GPR) framework produced significant improvements in both predictive performance and the modeling of residual spatial structure compared to its standalone components. This enhancement stems from the complementary strengths of the two methods: while the GAM component effectively captured the structured, non-linear effects of key covariates such as time, elevation, and precipitation, the GPR component successfully modeled the remaining spatial dependencies that the GAM could not account for. By combining these two approaches, the hybrid model provided a more comprehensive representation of the underlying data-generating process.

Quantitatively, the hybrid model outperformed both the standalone GAM and GPR models across all standard evaluation metrics, including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). These improvements indicate that the hybrid approach not only yielded more accurate predictions but also reduced unexplained variability, particularly in regions with complex spatial structure. The reduction in residual spatial autocorrelation further confirms that the GPR component effectively captured latent spatial patterns that the GAM alone was unable to address. The results are summarized in the table below:
The hybrid model achieved a pseudo-R² score of 0.9121, which indicates that it was able to explain over 91.2% of the variance in average temperature across the study area. This represents a significant improvement over the GAM (89.4%) and GPR (87.4%) models individually. The reduction in prediction error is similarly notable, with an RMSE of 1.28℃ and an MAE of 1.08℃, both of which were the lowest among the three approaches. These results confirm that the hybrid model effectively combines the strengths of its components: the structured predictability of GAM and the residual smoothing power of GPR.

A critical insight gained from the hybrid model is how effectively it addresses spatially structured residuals that persist after fitting the deterministic GAM component. While the GAM captured major nonlinear effects of elevation, precipitation, and location, residual analysis revealed lingering spatial autocorrelation – a common limitation in additive models that do not explicitly model correlation

Model	R^2 Score	RMSE	MAE	Mean Predictive Uncertainty
GAM	0.8939	1.47	1.23	—
GPR	0.8742	1.61	1.30	0.3475
Hybrid GAM–GPR	0.9121	1.28	1.08	0.3261

Table 1: Model's Predictive Performance

between neighboring spatial observations.

By training the GPR model on these residuals, the hybrid approach reconstructed the unexplained spatial structure using a Matérn covariance kernel. The Matérn kernel, with a smoothness parameter $\nu = 1.5$, offered a balance between model flexibility and spatial realism. Its learned length scales suggested that temperature observations remained correlated across distances of approximately 1.4° to 1.8°, corresponding to 150–200 kilometers — a typical spatial influence range for climatological processes.

This residual modeling allowed the hybrid model to adjust local temperature estimates based on surrounding residual patterns, improving prediction at both highly observed and under-sampled locations. The residual maps produced by GPR showed coherent, spatially smooth patterns that aligned with known microclimates and orographic zones. These improvements are particularly evident in topographically diverse areas such as the Rift Valley and Mount Kenya region, where traditional models often struggle to capture fine-scale variations.

Another important aspect of the hybrid model's performance lies in its probabilistic calibration. Unlike deterministic models, the GPR component provides posterior predictive variances at each location, offering spatially explicit uncertainty estimates. This capability is vital in decision-making processes related to agriculture, infrastructure, and disaster preparedness, where understanding prediction confidence is as important as the prediction itself.

In the hybrid model, the mean predictive standard deviation decreased to 0.3261 ℃, compared to 0.3475 ℃ in the standalone GPR model. This reduction reflects an increased certainty in predictions due to the GAM component capturing much of the explainable variance, thus leaving less uncertainty to be modeled by GPR. Moreover, uncertainty was spatially adaptive, – lowest in regions with dense observational coverage and consistent covariate behavior (e.g., central highlands) and higher in data-sparse or environmentally complex areas (e.g., northern arid zones and coastal regions).

This spatial distribution of uncertainty highlights the hybrid model's risk-aware predictive capability, allowing end-users to interpret not just what the model predicts but also how much trust can be placed in those predictions. For climate applications, this is particularly valuable in guiding the placement of weather stations, prioritizing regions for data acquisition, and designing robust policies that account for varying levels of climatic predictability.

3.2 Hybrid Model Validation

The hybrid GAM–GPR model produced highly accurate and reliable estimates of average temperature across the study area. Based on predictions generated on an independent test set, the model achieved a coefficient of determination (R^2) of 0.9121, indicating that it successfully explained 91.2% of the total variance in observed temperature values. This high level of explanatory power demonstrates the model's strong ability to generalize to unseen spatial locations.

In addition to its high R^2, the model yielded a Root Mean Squared Error (RMSE) of 1.28℃ and a Mean Absolute Error (MAE) of 1.08℃. These metrics confirm that the model consistently produced low prediction errors across the test set. The small difference between RMSE and MAE further suggests that the distribution of errors was fairly uniform, with no significant presence of large outliers.

The hybrid model also exhibited strong probabilistic calibration. The mean predictive standard deviation was 0.3261℃, indicating moderate and spatially adaptive confidence in the predictions. The lowest uncertainty values were observed in regions with dense data coverage, while areas with sparse observations, such as the northern arid zones, displayed predictably higher uncertainty. The magnitude and spatial structure of uncertainty were consistent with climatic intuition and observational density.

Visually, the predicted temperature surface from the hybrid model was smooth, continuous, and geographically coherent. High-temperature zones were clearly distinguished in the low-lying eastern and coastal regions, while cooler temperatures were consistently predicted in the elevated central and western highlands. Localized variations were effectively captured, particularly in regions with complex topography, such as the Rift Valley and around Lake Victoria. In these areas, the hybrid model generated fine-grained patterns that aligned with known microclimatic behavior.

Residual analysis confirmed the model's effectiveness in reducing unexplained spatial variability. The residuals were small and exhibited no strong spatial autocorrelation, indicating that the hybrid model captured both large-scale trends and local anomalies. Compared to the standalone GAM and GPR models, the hybrid approach yielded significantly reduced residual variance and tighter prediction intervals.

Overall, the results obtained demonstrate that the hybrid GAM–GPR model offers a robust and interpretable framework for temperature modeling. It balances accuracy with spatial adaptability and provides uncertainty-aware estimates that can support climate-sensitive decision-making and resource planning.

The results of the residual analysis and the Durbin-Watson test (2.03) strongly support the validity of the hybrid model. The residuals showed no evidence of autocorrelation or systematic error, and their distribution was consistent with the assumptions of regression and Gaussian process modeling. These diagnostics reinforce the robustness and reliability of the hybrid GAM–GPR framework, further justifying its use for climate modeling applications across heterogeneous spatial landscapes.

3.3 Hybrid Model Diagnostic Plots

Two diagnostic plots were generated to provide visual insight into the model's behavior: Residuals vs Actual Temperature and predicted vs. actual temperature, as shown in the figure below.

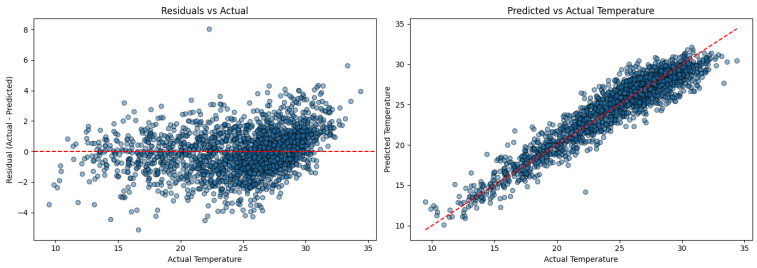

Figure 1: Hybrid Model Diagnostic Plots

Residuals vs Actual Temperature (Left Panel): This plot displays the residuals (i.e., actual minus predicted values) against the actual temperature observations. The red dashed line at zero serves as a reference line, where perfect predictions would lie. The residuals are fairly symmetrically distributed around the zero line, which suggests that the model does not exhibit systematic over- or underestimation across most of the temperature range.

While some minor dispersion is observed at higher temperatures (above 28°C), the residual spread remains within acceptable bounds, indicating that model error does not dramatically increase with temperature extremes. The residual pattern exhibits no distinct trends or curvature, supporting the assumption of residual randomness, a key requirement in validating the assumptions of hybrid modeling frameworks.

Predicted vs Actual Temperature (Right Panel): This plot compares the predicted temperature values against the actual observed temperatures. The ideal model would yield points lying exactly on the 45-degree reference line (red dashed line), indicating perfect agreement between predictions and observations. Most of the data points are tightly clustered around the line, indicating a strong linear relationship and high prediction fidelity.

The diagonal trend is preserved across the entire temperature spectrum (from $\sim$10°C to $\sim$35°C), suggesting that the model performs well across both low and high temperature regions. The tight clustering of points and absence of systematic deviation imply that the hybrid model provides stable, unbiased predictions across the range of observed values.

4 Conclusions

This study introduced and evaluated a hybrid modeling framework that integrates Generalized Additive Models (GAM) and Gaussian Process Regression (GPR) to predict average daily temperature across Kenya. The hybrid GAM–GPR model was designed to address key limitations observed in standalone statistical and machine learning models—namely, the lack of spatial correlation handling in GAMs and the limited interpretability of GPRs. By combining the structured, interpretable modeling capabilities of GAM with the spatially adaptive, uncertainty-aware predictions of GPR, the hybrid model provided a robust and flexible tool for spatio-temporal climate modeling.

Results demonstrated that the hybrid model outperformed its component models across all evaluated metrics, achieving higher accuracy ($R^2 = 0.9121$), lower prediction errors (RMSE = 1.28°C, MAE = 1.08°C), and better residual behavior. Moreover, the model offered spatially explicit uncertainty estimates, a critical feature for decision-making in climate-sensitive sectors. The

diagnostic plots and residual analyses confirmed the hybrid model's ability to capture both global trends and localized spatial patterns, especially in climatically and topographically complex areas such as the Rift Valley and central highlands. Overall, the hybrid GPR–GAM framework provides a scientifically grounded and operationally practical solution for localized climate prediction, with strong potential for integration into climate resilience planning, agricultural advisories, infrastructure design, and environmental risk assessment in Kenya and similar geographies.

While the hybrid GAM–GPR model has demonstrated strong predictive capabilities, several directions remain for extending and refining this work. First, future research could explore *multi-output modeling* to jointly predict multiple climate variables—such as temperature, precipitation, and humidity—within a unified framework. This would provide a more holistic understanding of climate dynamics and improve integrated risk modeling. Second, the *scalability* of GPR remains a challenge for very large datasets; thus, incorporating *sparse Gaussian processes* or *variational inference techniques* could significantly reduce computational overhead while maintaining predictive accuracy.

Third, the model could be extended to include additional covariates such as *Normalized Difference Vegetation Index (NDVI)*, land use, or soil moisture data, which may enhance the model's capacity to represent land-atmosphere interactions. Another promising direction involves *real-time forecasting*, where the model is adapted for streaming data and integrated into early warning systems for drought or heatwaves. Finally, the applicability of the hybrid framework should be tested in *other geographical regions* with diverse climatic and topographic conditions, to evaluate its generalizability and to inform broader climate adaptation strategies across the Global South. These enhancements would not only increase the model's practical utility but also contribute significantly to the growing field of interpretable and uncertainty-aware climate modeling.

COMPETING INTERESTS DISCLAIMER:
Authors have declared that they have no known competing financial interests OR non-financial interests OR personal relationships that could have appeared to influence the work reported in this paper.

References

P. K. Thornton, P. J. Ericksen, M. Herrero, and A. J. Challinor, "Climate variability and vulnerability to climate change: a review," *Global change biology*, vol. 20, no. 11, pp. 3313–3328, 2014.

M. Herrero, C. Ringler, J. V. D. Steeg, P. K. Thornton, T. Zhu, E. Bryan, A. Omolo, J. Koo, and A. M. O. Notenbaert, "Climate variability and climate change and their impacts on kenya's agricultural sector," 2010.

T. J. Lawrence, J. M. Vilbig, G. Kangogo, E. M. F`evre, S. L. Deem, I. Gluecks, V. Sagan, and E. Shacham, "Shifting climate zones and expanding tropical and arid climate regions across kenya (1980–2020)," *Regional Environmental Change*, vol. 23, no. 2, p. 59, 2023.

N. Chokkavarapu and V. R. Mandla, "Comparative study of gcms, rcms, downscaling and hydrological models: a review toward future climate change impact estimation," *SN Applied Sciences*, vol. 1, no. 12, p. 1698, 2019.

A. Pathirana, H. B. Denekew, W. Veerbeek, C. Zevenbergen, and A. T. Banda, "Impact of urban growth-driven landuse change on microclimate and extreme precipitation—a sensitivity study," *Atmospheric Research*, vol. 138, pp. 59–72, 2014.

A. Wilson and M. R. Anwar, "The future of adaptive machine learning algorithms in high-dimensional data processing," *International Transactions on Artificial Intelligence*, vol. 3, no. 1, pp. 97–107, 2024.

M. Wiley, J. F. Wiley, M. Wiley, and J. F. Wiley, "Gams," *Advanced R Statistical Programming and Data Models: Analysis, Machine Learning, and Visualization*, pp. 165–224, 2019.

P. Zschech, S. Weinzierl, N. Hambauer, S. Zilker, and M. Kraus, "Gam (e) changer or not? an evaluation of interpretable machine learning models based on additive model constraints," *arXiv preprint arXiv:2204.09123*, 2022.

L. Pipia, E. Amin, S. Belda, M. Salinero-Delgado, and J. Verrelst, "Green lai mapping and cloud gap-filling using gaussian process regression in google earth engine," *Remote Sensing*, vol. 13, no. 3, p. 403, 2021.

D. Tuia, M. Volpi, J. Verrelst, and G. Camps-Valls, "Advances in kernel machines for image classification and biophysical parameter retrieval," *Mathematical Models for Remote Sensing Image Processing: Models and Methods for the Analysis of 2D Satellite and Aerial Images*, pp. 399–441, 2018.