Urban Heat and Health in Johannesburg: A Multidimensional Analysis of Vulnerability, Explanatory Modelling, and Predictive Outcomes

PhD proposal of Mr Craig Parker **(MSc, MM-PDM, BS)**

30/05/2024 version 2.1

Table of Contents

[1. Summary 3](#_Toc167976196)

[2. Introduction 4](#_Toc167976197)

[3. Study setting 5](#_Toc167976198)

[4. Aims and objectives 6](#_Toc167976199)

[5. Data description 6](#_Toc167976200)

[Socio-economic and environmental data 7](#_Toc167976201)

[Health trials and cohort data 7](#_Toc167976202)

[Integration of datasets 9](#_Toc167976203)

[Managing bias 10](#_Toc167976204)

[6. Methods 10](#_Toc167976205)

[Quantifying intra-urban socio-economic and environmental heat vulnerability 11](#_Toc167976206)

[Delineating time-lagged, non-linear heat-health dynamics through explanatory machine learning models 12](#_Toc167976207)

[Developing a Spatially and Demographically Stratified Heat-Health Outcome Model 13](#_Toc167976208)

[7. Ethical Considerations: 14](#_Toc167976209)

[8. Work Plan: 14](#_Toc167976210)

[9. Research Outputs 15](#_Toc167976211)

[10. POPIA compliance and protection of personal information 16](#_Toc167976212)

[11. Strengths and Weaknesses 16](#_Toc167976213)

[12. Budget 17](#_Toc167976214)

[13. Supervisors 17](#_Toc167976215)

[14. Conclusion 18](#_Toc167976216)

[15. References 19](#_Toc167976217)

## Summary

This research proposal aims to investigate the intricate relationship between urban heat and health in Johannesburg, focusing on the socio-economic and environmental factors that contribute to heat vulnerability. The study is structured around three main objectives:

1. To map intra-urban heat vulnerability using socio-economic data combined with spatial information.
2. To understand the complex non-linear and time-lagged relationships between environmental factors and health outcomes.
3. To develop a predictive model for heat-related health outcomes.

The methodology includes three key methods to achieve these objectives:

1. **Mapping Urban Heat Vulnerability:** We will combine socio-economic data with spatial information from sources like Copernicus ERA5 reanalysis and Landsat imagery. Principal Component Analysis (PCA) will be used to identify key predictors, followed by multi-level clustering to categorize urban areas based on socio-economic and environmental data.
2. **Explanatory Modeling of Heat-Health Relationships:** Advanced machine learning techniques such as Random Forests and XGBoost will be employed to understand the complex non-linear and time-lagged relationships between environmental factors and health outcomes. This phase is crucial for elucidating the underlying mechanisms of heat-health interactions in Johannesburg.
3. **Predictive Heat-Health Outcome Modelling:** Building on insights from the initial stages, we will develop a predictive model for heat-related health outcomes. This model will focus on identifying socio-economic and environmental conditions that heighten health risks and will employ supervised machine learning techniques for predictive accuracy. The predictive model aims to stratify risk across different demographic groups, aiding in targeted health interventions and resource planning.

Johannesburg, the study setting, exhibits a subtropical highland climate(Köppen classification as Cwb), socio-economic disparities, and health landscape. The research will utilise various data sources, including socio-economic and environmental data, health trials and cohort data. Integrating these datasets will allow for a comprehensive re-analysis of heat's impact on health in Johannesburg.

Ethical considerations will be addressed, focusing on informed consent for secondary data usage and protecting potentially identifiable information. The research will adhere to regulations for protecting human subjects and comply with the Protection of Personal Information Act of South Africa (POPIA).

The work plan outlines a four-year timeline, with activities ranging from a comprehensive literature review and data analysis to developing and refining the predictive model. The research outputs will include four seminal papers, each highlighting a key aspect of investigating heat-health outcomes in Johannesburg.

Overall, this research aims to deepen our understanding of heat-health interactions in large African cities and provide actionable insights for public health strategies and interventions in the context of climate change.

## Introduction

High ambient temperatures above long-term averages during summer months and discrete heat extremes (e.g. heat waves) are associated with excess mortality and considerable morbidity[1-4]. Climate projections over sub-Saharan Africa show that heatwaves will continue to become hotter and more dangerous, even if global warming is [kept below 1.5C](https://www.carbonbrief.org/in-depth-qa-ipccs-special-report-on-climate-change-at-one-point-five-c#impacts). Combined with population changes, [20-50 times as many people](https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2018EF001020) could be exposed to dangerous heat in African cities by the end of the century[5]. Anthropogenic climate change has resulted in a more than 1.2°C rise in temperature globally since the Industrial Revolution.[6]. This increase is not evenly distributed across the planet, however, or even within local areas[7]. Regional differences and the effect of urban development and land use change mean that many parts of Africa are experiencing higher-than-average temperature increases and more frequent, intensive and longer-lasting heat waves[8]

This PhD proposal, emanating from the NIH funded HE²AT Center Research Project (RP 2), aims to delve into the complex interplay between urban environments and their impact on heat-health, spotlighting the urgent need for nuanced responses[9]. It foregrounds the acute risks faced by vulnerable populations in urban settings—including the economically disadvantaged, the elderly, those with pre-existing health conditions such as TB and lung disease, children, outdoor workers, and residents of densely populated or informal settlements[10-15] [16-18] [10, 14, 19-22]s. For these groups, the urban heat island (UHI) effect is not just a concept but a daily reality that worsens existing vulnerabilities[23-25].

Amidst a backdrop of climate change, urbanising African cities have become hotspots for exacerbated public health risks. These risks are amplified by the UHI effect, where urban development and sparse vegetation lead to significantly higher temperatures within cities than in their surrounding rural areas[26] [7] . This scenario is particularly alarming in areas where housing and infrastructure, often comprising informal dwellings and low-cost housing, fail to provide adequate thermal comfort, leading to indoor temperatures that are notably higher than the external environment[27, 28]. Such conditions intensify heat exposure and highlight the inadequacies in current urban planning and housing design, underscoring the imperative for localised interventions.

Building on foundational research, including significant studies from Johannesburg, this project addresses the highlighted gap by proposing a comprehensive examination of heat-health risks and their socio-economic and infrastructural determinants in Africa. The critical works of Ncongwane et al. (2021), Pasquini et al. (2020) and Wright et al. (2019), alongside investigations into nighttime heatwaves by Eghosa Igun et al. (2022) and health impacts assessments by Enete et al. (2017), underscore the complexity of these challenges[29-31]. These studies collectively point to an escalating threat of heatwaves, intensified by UHI effects and the resultant health burdens, particularly in cities. [32].

The research further acknowledges the compounded vulnerabilities in African urban environments—marked by high disease prevalence rates, scarce cooling resources, and extensive informal settlements—which heighten the health risks associated with rising temperatures[33, 34] [35]. Despite the acknowledged impacts of heat on health, a knowledge gap persists in effectively assessing and predicting heat-related health risks within African urban contexts. Current assessments often overlook the multifaceted nature of urban spaces, unique environmental exposures, and the specific demographic and disease profiles prevalent in the continent.

By proposing to deepen the understanding of heat-related health risks in urban African settings, this research aims to significantly contribute to the field of urban public health amid climate change. It seeks to advance our knowledge and capabilities in predicting and mitigating heat-health challenges, aiming to foster more resilient urban environments across Africa. This endeavour responds to an urgent public health need and lays the groundwork for future research and policy-making aimed at protecting the most vulnerable populations from the adverse effects of urban heat.

## Study setting

Nestled on the South African Highveld plateau at an elevation of 1,700 meters, the vibrant city of Johannesburg forms the setting of this research. As the largest city in South Africa and the 26th largest globally, Johannesburg's population exceeds 5.635 million inhabitants[36]. This bustling metropolis, characterised by its unique subtropical highland climate(Köppen classification as Cwb), provides a compelling backdrop for exploring urban heat health impacts[37].

Johannesburg's distinct weather patterns follow a bifurcated climate cycle. Summer months, extending from October to April, are marked by hot days often followed by refreshing afternoon thundershowers, transitioning into cooler evenings. The winter period from May to September offers a contrasting dry, sunny day leading into cold nights. Due to the city's high elevation, the climate remains generally mild, with average maximum daytime temperatures oscillating between 25.6 °C (78.1°F) in January and 16 °C (61°F) in June.[38].

Heatwaves pose a significant public health risk in Johannesburg.[37]. Research has shown that passing temperature and humidity thresholds are associated with a heightened risk of mortality and morbidity. Such insights emphasise characterising past and future heat waves to enhance heat-health warning systems and inform health-centric policy-making.[39].

Central to understanding the impact of urban heat health in Johannesburg is the consideration of social determinants of health (SDoH). These determinants include factors such as economic stability, access to educational opportunities, healthcare services, and the quality of housing and its design, all of which profoundly influence public health outcomes. In Johannesburg, the stark economic disparities manifest in “Green Apartheid”, where varying levels of access to critical resources are significantly shaping the residents' capacity to adapt to and mitigate the effects of urban heat[40]. The interplay between these social determinants and urban heat exposure underscores the complexity of health vulnerabilities in the city, highlighting the need for interventions that address not only the environmental aspects of heat, but also the underlying social inequities.[41].

Johannesburg's socio-economic canvas is marked by stark disparities, with impoverished urban communities shouldering the disproportionate burden of climate change impacts on health and well-being[42]. Inadequate housing, limited access to resources, and poverty elevate communities' vulnerability to heat-related health effects, a situation worsened by infrastructural deficiencies[37, 43].

From a health standpoint, a unique set of risk factors shapes the relationship between heat and health in Johannesburg. Adverse health outcomes linked to heat exposure include high blood pressure, respiratory stress, and cardiac conditions, further aggravated by the prevailing socio-economic and infrastructural conditions[23, 31, 44]. A critical health consideration in Johannesburg is the high prevalence of communicable diseases, notably HIV, tuberculosis, and COVID-19 [45, 46]. These diseases add complexity to the health landscape, with heat exposure potentially affecting the health status and disease progression in affected individuals.

## Aims and objectives

The primary objective of this research study is to deepen our understanding of the complex, spatially, and demographically stratified heat-health interactions that occur daily in Johannesburg. This is driven by the urgent need to address the public health challenges posed by increasing urban heat, particularly in the context of climate change and socio-economic disparities.

* **To identify and map intra-urban heat vulnerability and exposure across Johannesburg:** This is essential to highlight areas and populations that are most at risk. Understanding where vulnerability is highest can help target interventions and allocate resources more effectively to protect those who are most vulnerable.
* **To employ machine learning explanatory models:** By uncovering and interpreting the intricate relationships between climate variables and health outcomes, this aim seeks to enhance our understanding of heat-health dynamics. This focus on the explainability and interpretability of model findings is crucial for developing scientifically sound and practically applicable strategies for policymakers and public health officials.
* **To develop a spatially and demographically stratified heat-health outcome prediction model:** This model will predict the probability of adverse health outcomes at different temperature thresholds. The goal is to provide a robust tool for forecasting and mitigating health risks, enabling more precise and effective public health responses to heat events.

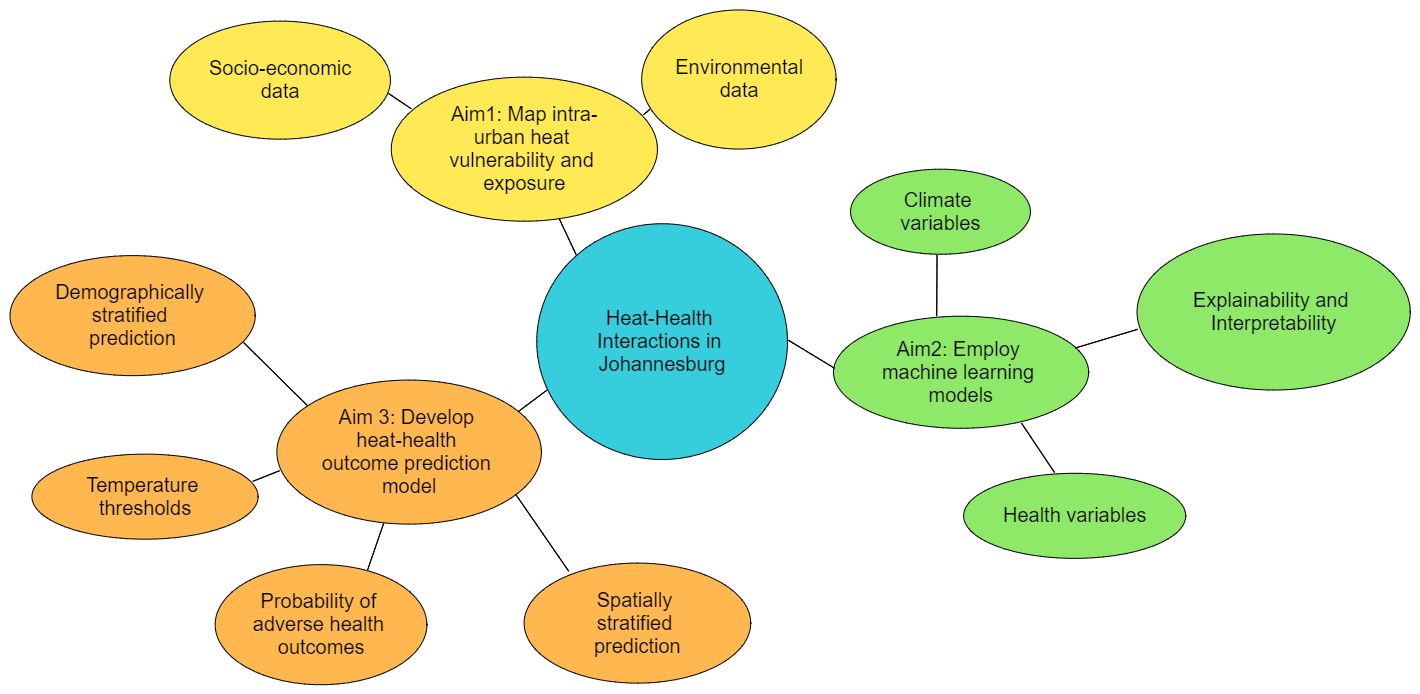


Figure 1: Summary of study aims.

## Data description

This research will utilise a wide array of data sources to achieve its objectives. Data will be housed and managed within the high-performance computing facilities at the University of Cape Town's Climate System Analysis Group (UCT CSAG), a central asset developed in the HE2AT centre project. Details of the data management plan that oversees this system is included in Annex: A.

### Socio-economic and environmental data

This research will collect socio-economic geospatial data, which includes information on household economic conditions, service availability, and residential characteristics—referring to factors like housing type, construction materials used, and the quality and condition of living spaces [47]. The data will include national census records, specialised household, and demographic surveys and encompass details about individual and household income, education, occupation, living circumstances, and accessibility to healthcare, education, and transportation services. [48] The Gauteng City-Region Observatory (GCRO) datasets will provide many of these key variables. [48, 49].

Remote sensing data will be retrieved from satellite sensors, including optical images and indicators of physical aspects such as land surface temperature, soil moisture, vegetation condition, and land use [50]. Where available, researchers will amalgamate data from current sensor networks with urban land use and building density details to create a model of urban land use heat [47, 48]. Although Landsat and MODIS data primarily measure land surface temperature (LST), statistical models can estimate air temperature from remotely sensed LST. However, it should be noted that LST may not fully capture heat stress experienced in urban areas. In this study, appropriate statistical models will be used to indirectly retrieve air temperature from the LST data provided by Landsat and MODIS, and where possible, we will incorporate humidity data to provide a more comprehensive assessment of heat stress [51].

Climate-associated data will be sourced from open data repositories, such as the Copernicus Climate Data Store (CDS) and Earth System Grid Federation (ESGF), offering observational-based datasets, historical re-analyses, and climate simulations. While the Copernicus Climate Data Store (CDS) and Earth System Grid Federation (ESGF) provide valuable climate data, their spatial resolution may not be sufficient to distinguish different parts of the city[52]. To address this limitation, we will employ downscaling techniques to enhance the spatial detail of our geospatial climate data. Specifically, we will explore the use of dynamic downscaling with high-resolution climate models such as the Weather Research and Forecasting (WRF) model and the UrbClim urban climate model. These models offer detailed results on heat stress for cities, allowing for a more precise analysis of intra-urban heat variations and can improve the accuracy of our heat risk assessments for Johannesburg [53, 54].

Additionally, the IBM-PAIRS platform will be employed as a source of climate data, including data from climate models, weather stations, and satellite observations[55]. To further enhance our analysis, we will integrate datasets from the European Space Agency's WorldCover portal and the Global Human Settlement Layer (GHSL), which provide detailed land cover and human settlement data, respectively[56, 57].This will provide a comprehensive snapshot of Johannesburg’s past and future climate and population conditions, including the frequency, duration, and intensity of heat waves with a growing population.

### Health trials and cohort data

The health data for this study will be collected from clinical trials and cohort studies, such as HIV drug trials and COVID-19 vaccine trials. These studies typically involve hundreds to thousands of participants and are conducted over an extended period (multiple years) within a specific geographical area. They provide detailed longitudinal individual health data for building machine-learning models relating time-varying predictors to health outcomes. Potential outcomes of interest include cardiovascular events, respiratory issues, kidney conditions, and mental health impacts, which may be exacerbated by heat exposure in urban environments[58].

More specifically, the health cohort data integrated into the study will be identified based on the availability of three classes of variables within each study:

* Clinical variables: including vital signs (e.g., body temperature, blood pressure, and heart rate), indicators of heat-related illness (e.g., headache, dizziness, fatigue, and nausea), and details on pre-existing medical conditions (e.g., hypertension, diabetes, and cardiovascular disease) that could increase the risk of heat-related illness, and documentation of adverse events potentially related to heat exposure.
* Laboratory variables: including blood tests (e.g., electrolyte levels, liver and kidney function tests), markers of inflammation and oxidative stress, as well as HIV tests, including viral load and CD4 count, and COVID-19 test results.
* Demographic and SDOH variables: involving basic demographic information (e.g., age, sex, race, and ethnicity), socio-economic factors (e.g., education, income, and occupation), and data on housing and urban infrastructure (e.g., air conditioning availability, ventilation, and shading) that could influence heat exposure and the degree to which individuals and households are at an increased risk.

In response to the shifts in mortality and morbidity during the 2020-2022 COVID-19 pandemic, we will analyse data separately for pre-pandemic, pandemic, and post-pandemic periods. Additionally, we will include COVID-19-related variables as covariates in our models to control for the pandemic's impact on health outcomes.

Table 1: Summary of Data Sources for each Objective

|  |  |
| --- | --- |
| **Objective** | **Data Sources** |
| 1. **Map intra-urban heat vulnerability and exposure** | - Socio-economic data (census, surveys, GCRO datasets)  - Geospatial data (land use, building density, OpenStreetMaps)  - Climate data (WRF, UrbClim models, downscaled CDS & ESGF data, IBM-PAIRS platform) |
| 2. **Employ machine learning explanatory models to uncover and interpret the intricate relationships between climate variables and health outcomes** | - Health data with clinical variables (e.g., vital signs, heat-related illness indicators)  - High-resolution urban temperature hazard maps (Landsat, MODIS data with statistical models for air temperature estimation)  - Remote sensing data (satellite imagery, land surface temperature, soil moisture, vegetation condition)  - Socio-economic and environmental data (household economic conditions, service availability, residential characteristics) |
| 3. **Develop a spatially and demographically stratified heat-health outcome prediction model** | - Integrated health and socio-economic data  - Geospatial heat hazard maps  - Health outcome forecast model outputs  - COVID-19 incidence and mortality rates (for pandemic period adjustment)  - Risk profile data (demographic groups, health conditions, locations, socio-economic statuses) |

### Integration of datasets

A critical component of this research is the integration of diverse datasets, including socio-economic, clinical, environmental, and geospatial data, to provide a comprehensive analysis of heat's impact on health in African cities. The integration process will begin with data preparation, where each dataset will undergo preprocessing to ensure consistency in formats, units of measurement, and timeframes. This step will involve data cleaning, normalization, and standardization to facilitate seamless integration.

Following data preparation, we will align all datasets to a common geospatial framework, a process that involves geocoding health trial participant locations and aligning them with socio-economic and environmental data layers. To protect privacy while retaining spatial trends, we will apply spatial jittering techniques to the health trial participant geolocations. Temporal synchronization of datasets is also crucial, especially for analyzing time-lagged effects of heat exposure on health outcomes. We will ensure that all datasets are aligned to the same temporal resolution, allowing for accurate temporal analysis.

The data fusion step will involve combining datasets using common identifiers such as geographic coordinates and time stamps. We will employ data fusion techniques that allow for the combination of datasets with different spatial and temporal resolutions, ensuring that the integrated dataset maintains the highest possible level of detail. Post-integration, we will conduct quality checks to ensure the integrity of the integrated dataset. This includes verifying the accuracy of merged data, checking for any anomalies introduced during the integration process, and ensuring that the dataset adequately represents the study area.

Finally, the integrated dataset will be stored in a secure, accessible database within the high-performance computing facilities at the University of Cape Town's Climate System Analysis Group (UCT CSAG). Appropriate access controls will be implemented to ensure that the data is accessible to authorized researchers while maintaining confidentiality and compliance with data protection regulations. By following these steps, we aim to create a robust, integrated dataset that provides a holistic view of the interplay between heat exposure, socio-economic factors, and health outcomes in Johannesburg. This integrated dataset will serve as the foundation for our subsequent analyses and modelling efforts.

**Table 1: Eligibility Criteria for Research Project 2**

|  |  |
| --- | --- |
| Criteria | Description |
| Study type | Cohort or trial with at least 200 adult participants |
| Study location | Johannesburg or Abidjan, or both cities |
| Study design | Randomised or non-randomised clinical trial, or observational or interventional cohort with prospectively collected data |
| Data collected | At least two of the clinical or lab variables |
| Ethics approval | Local ethics approvals obtained |

For the success of this project, access to relevant trials and cohort data is crucial. In the event of data unavailability or sharing restrictions, we have contingency plans to ensure the project's progression. These include exploring alternative data sources such as the National Health Laboratory Service (NHLS), adjusting the study's scope, and utilising synthetic data if necessary.

### Managing bias

In our study, managing potential biases is paramount to ensure the integrity and robustness of our findings. Our primary strategy for managing bias involves the careful selection of health data sources. We will ensure that these sources meet established quality criteria and represent diverse demographic and geographic segments within Johannesburg. This approach will help us avoid selection bias that could skew our findings.

When analysing the data, we will employ specific statistical methods to adjust for potential biases. These methods include propensity score matching, inverse probability weighting, and stratification. Propensity score matching will be used to create comparable groups based on observed covariates, reducing selection bias. Inverse probability weighting will adjust for the probability of treatment assignment, controlling for confounding variables. Stratification will involve dividing the data into subgroups based on key variables, allowing us to assess the effect of heat exposure within each stratum.

Furthermore, we will conduct sensitivity analyses to assess the robustness of our findings. This involves examining how variations in the data and assumptions affect the results. By doing so, we can confirm the stability of our findings and identify any potential biases that may influence the results.

By implementing these strategies, we aim to minimize bias in our study and ensure that our findings accurately reflect the impact of heat exposure on health outcomes in Johannesburg.[59].

## Methods

A diagram of a company

Description automatically generated

Figure 2:Methodological Framework for Predictive Heat-Health Outcomes in Johannesburg

### Quantifying intra-urban socio-economic and environmental heat vulnerability

The methodology for quantifying intra-urban socio-economic and environmental vulnerability to heat in Johannesburg will commence with a collaborative causal mapping phase. This initial stage will involve experts from the Heat Center, including health researchers, social scientists, climate scientists, policymakers, the Department of Health, and representatives from the general public. Through workshops and sessions, this diverse group will engage in mapping processes to identify and interconnect variables contributing to heat vulnerability, aiming to develop a comprehensive causal loop diagram specific to Johannesburg's urban landscape.

Once the causal loop diagram is established, we will employ Principal Component Analysis (PCA) to reduce dimensionality within the collected data. The PCA will be pivotal in identifying variables that significantly impact heat vulnerability. We anticipate environmental measures such as Urban Thermal Field Variance Index (UTFVI), Land Surface Temperature (LST), and Normalized Difference Vegetation Index (NDVI), along with health metrics like chronic disease prevalence rates, will emerge as critical contributors. These variables will be scrutinized for their roles in shaping the urban heat vulnerability landscape.

Following PCA, we will synthesize the extracted principal components into a composite vulnerability index. This index aims to capture the aggregate socio-economic and environmental susceptibilities, with a focus on heat vulnerability as the primary climatic threat. The construction of this index will involve selecting variables based on their contribution to the principal components, ensuring that the most influential factors are included.

Subsequently, the methodology will involve a spatial multi-criteria analysis, whereby the principal components and composite index inform the generation of a vulnerability map. This map will delineate areas within Johannesburg at the greatest vulnerability, serving as an essential guide for directing policy interventions and resource allocation towards increasing urban resilience against heat. The multi-criteria analysis will be conducted using a weighted overlay approach, where each variable in the composite index is assigned a weight based on its relative importance, as determined by the PCA. These weights will then be used to combine the variables, producing a final vulnerability score for each spatial unit.

The proposed methodology, from collaborative mapping to vulnerability map production, will offer a sophisticated framework to advance our understanding of urban heat vulnerability. It promises to deliver actionable insights, empowering stakeholders to enact informed and targeted strategies for mitigating the impacts of urban heat.

A diagram of a diagram

Description automatically generated

Figure 3: Proposed causal mapping example of Heat Vulnerability framework in Johannesburg

### Delineating time-lagged, non-linear heat-health dynamics through explanatory machine learning models

Our approach leverages the strengths of explanatory machine learning models, specifically Random Forests and XGBoost, to dissect and understand the time-lagged, non-linear interactions between socio-economic and environmental vulnerabilities and health outcomes in Johannesburg. These models are chosen for their interpretability, precision in handling diverse data types and structures, and feature selection capabilities [60, 61]

Random Forests, known for their interpretability and ability to rank features based on their importance, provide a robust framework for capturing the most significant predictors without explicit variable exclusion. The model's output on feature importance will guide our understanding of the key determinants and their respective influence on health outcomes[62, 63].

On the other hand, XGBoost, with its precision in handling diverse data types and structures, will be employed to detail the nuances of the data. Its feature selection and regularisation capability makes it an optimal choice for identifying and interpreting critical features. We will use XGBoost's SHAP values to interpret the contribution of each feature within the context of time-lagged effects, thus emphasising the explanatory aspect of our analysis.

In adherence to best practices, the entire dataset will be utilised for our explanatory models to allow the machine learning algorithms to internally assess the importance of each feature without withholding any portion of the data for hold-out validation. This approach ensures that our interpretation of the model is based on the complete information available, providing a comprehensive view of the heat health dynamics at play[64].

Our commitment to methodological rigour involves performing a sensitivity analysis to validate the consistency and reliability of the feature importance outcomes. By examining how variations in the data affect the model results, we can confirm the stability of our explanatory factors.

Additionally, we will employ bootstrapping methods to assess the stability of our feature importance rankings and the robustness of the model's predictive power under various data sampling scenarios. The statistical significance of the model-derived relationships will be evaluated using permutation tests, which will allow us to discern the predictive power of features from chance associations.

The interpretative power of machine learning will be harnessed to its fullest to uncover the temporal and complex associations within our urban health data, offering clear insights into the interactions between the environment, time, and health. This will enable stakeholders to grasp the multifaceted nature of heat-health vulnerabilities and craft targeted interventions informed by a thorough understanding of the determinants.

Through this focused and methodologically robust approach, we aim to provide a transparent and detailed explanation of the factors that contribute to heat-related health risks, contributing significantly to urban public health research.

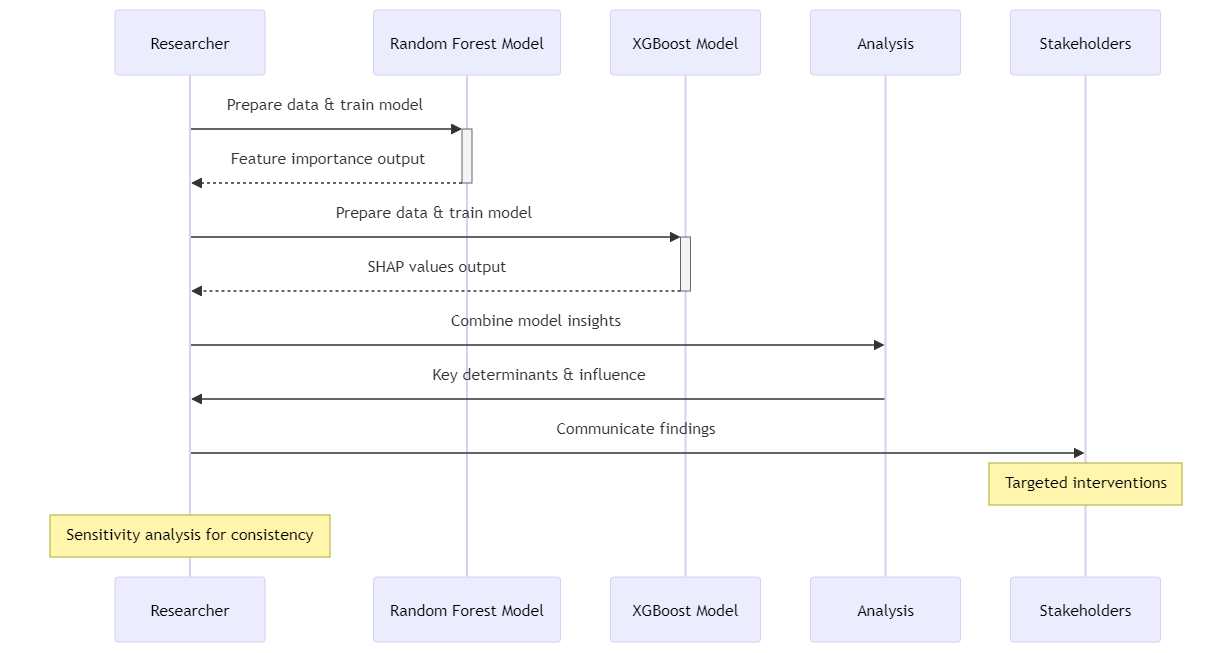


Figure 4: Methodological outline of explanatory ML modelling

### Developing a Spatially and Demographically Stratified Heat-Health Outcome Model

In this research, we aim to develop a predictive model that forecasts heat-related health outcomes at the individual or ward level in Johannesburg by integrating socio-economic, environmental, and health data. The process begins with data integration and preprocessing, where each record in the unified dataset represents an individual or a ward, with features including demographic information, environmental factors, and health indicators. The data will be preprocessed to handle missing values, normalize numerical features, and encode categorical variables.

Feature selection techniques such as mutual information, correlation analysis, and recursive feature elimination will be employed to identify the most relevant features for predicting heat-related health outcomes, reducing the dimensionality of the dataset and improving the model's performance. We will experiment with various machine learning algorithms, including decision trees, random forests, gradient boosting machines, and neural networks, to find the best-performing model based on the nature of the data and the desired balance between interpretability and predictive accuracy.

The model will be trained and validated using cross-validation techniques, with performance evaluated using metrics such as accuracy, precision, recall, and area under the ROC curve (AUC-ROC). To stratify risk across different demographic groups, the model will output a risk score for each individual or ward, reflecting the likelihood of experiencing heat-related health issues based on the input features. This stratification will allow for targeted interventions and resource allocation to high-risk groups, such as elderly individuals, low-income households, or wards with limited green space.

Model interpretation will be facilitated using techniques such as SHAP values to understand the contribution of each feature to the overall risk score, ensuring that the findings can be communicated effectively to policymakers and healthcare providers. Once validated, the predictive model will be deployed as a tool for public health officials and urban planners to assess and mitigate heat-related health risks in Johannesburg, with regular updates to ensure its continued relevance and accuracy.

## Ethical Considerations

This research study received ethical approval from both the Wits Human Research Ethics Committee in Johannesburg (reference number 220606) on June 30, 2022) and will follow the United States Department of Health and Human Services regulations for the protection of human subjects in research (45 CFR 46). Our research protocol has two critical ethical and legal considerations: informed consent for secondary data usage and the protection of potentially identifiable information.

Regarding informed consent for secondary data usage, we will critically examine the consent procedures intended for the original study. If a participant has previously provided "broad consent", permitting the use of their data in future research endeavours, we can share their data without additional ethical approvals. For participants who have granted "narrow consent, " which restricts data sharing beyond the original study purpose, careful deliberation is required. If obtaining renewed consent is unfeasible or involves a disproportionate effort, we will seek an informed consent waiver from the appropriate ethics committee.

In order to protect potentially identifiable information and minimising privacy risks (such as indirect identifiers like geographical data in the collected data) we will employ several protective measures including the restriction of identifiable data and no use of real names or other identifying factors. Data will be stored in a password-protected server with limited access. Additionally, following data minimisation principles, we will retain only the data essential for achieving our study objectives. When applicable, we will anonymise data through geographical aggregation and jittering, especially when home addresses are used.

Finally, we acknowledge the specific legislative requirements for using health data in different countries, including the laws surrounding the cross-border transfer of such data. We will, therefore, require data providers to provide a contractual guarantee, as part of the data sharing agreement, that all original studies followed appropriate informed consent procedures and that the sharing of this data complies with all relevant data protection laws.

## Work Plan

|  |  |  |
| --- | --- | --- |
| Year | Activities | Outcomes |
| 1 | Conduct a comprehensive literature review. Establish research protocol. The draft first paper on intra-urban heat vulnerability. | First paper draft on intra-urban heat vulnerability. |
| 2 | Analyse GCRO and climate data. Inform the second paper on intra-urban socio-economic and environmental vulnerability. | Second paper on socio-economic and environmental vulnerability. |
| 3 | Apply advanced machine learning techniques. Investigate lagged impacts of heat-health exposures. Draft third paper on heat and health associations. | The third paper on lagged impacts of heat-health exposures. |
| 4 | Develop and refine a predictive model. Evaluate model performance. Document model development in the fourth paper. Complete PhD thesis. | Fourth paper on the predictive model. Completed PhD thesis |

## Research Outputs

Our research endeavors will culminate in the publication of four seminal papers, each highlighting a key facet of our investigation into heat-health outcomes in Johannesburg. These papers are pivotal to our academic contribution and will be disseminated widely for maximal impact.

1. **Research Protocol Documentation Paper**: This initial paper will outline the comprehensive research protocol used in our study. It will detail the methodological framework, aiming to provide a replicable model for similar studies. This paper's public availability will not only validate our scientific approach but also encourage further research in this vital domain.
2. **Socio-economic and Climate Vulnerability Analysis Paper**: The second paper will delve into the socio-economic and climate data analysis, focusing on identifying the vulnerability traits within the Johannesburg population. We plan to present these insights at scientific conferences and publish in open-access journals, stimulating discussions that extend beyond the academic sphere and contribute to a broader understanding of the socio-economic impacts of climate change.
3. **Heat-Health Correlations and Explanatory ML Modelling Analysis Paper**: The third paper will showcase the results of using advanced statistical and machine learning explanatory models to analyze the complex relationships between temperature fluctuations and health outcomes, particularly focusing on the time-lagged impacts of heat exposure in Johannesburg. This paper will highlight the efficacy of explanatory ML models in unraveling these intricate relationships, providing valuable insights that can inform future research directions and public health policy decisions.
4. **Heat-Health Outcome Prediction Model Paper**: The final paper will focus on the development and validation of our heat-health outcome prediction model. It will detail the model's efficacy in forecasting health risks and its potential to guide risk mitigation strategies. By sharing this model, we seek to foster proactive, data-driven public health initiatives, both locally in Johannesburg and in similar urban contexts globally.

These papers will form the cornerstone of our scientific communication and outreach, underpinning presentations at academic forums and engagements with community and policy stakeholders. They are intended to significantly contribute to the scholarly dialogue on climate and health, while also informing public policy, raising awareness, and guiding future adaptation strategies in the face of climate change

## POPIA compliance and protection of personal information

Our research meticulously attends to data security and confidentiality in alignment with the Protection of Personal Information Act of South Africa (POPIA, 2013). POPIA limits personal information processing but allows its use in scientific research. Our study is cognizant of this, alongside other governing legal frameworks like the National Health Act No 61 of 2003, the Constitution of the Republic of South Africa, and the Department of Health guidelines on Ethics in Health Research.

Our research strategy includes processing de-identified health databases in which re-identification is virtually impossible. Where personal information has not been de-identified, we comply with the relevant sections of POPIA, allowing us to process health data for historical, statistical, and research purposes.

The information gathered and processed by our team will only be used for research and statistical purposes, which directly relate to addressing the major public interest of understanding and mitigating the health implications of rapidly escalating temperatures and heat waves, particularly in Africa. This processing of data is deemed necessary and justified as it serves to inform strategies to combat one of the greatest health threats of the 21st century – climate change.

Security measures will be implemented to prevent unlawful access or processing of personal information, while the operators involved in the data handling process will be bound by a written contract, ensuring accountability. This approach aligns with Sections 19, 20, and 21 of POPIA, demonstrating our commitment to preserving the rights of individuals and upholding the highest ethical standards in scientific research.

## Strengths and Weaknesses

Study strengths and limitations

1. Employs comprehensive data collection from clinical, socio-economic, and remote sensing sources, ensuring a multidimensional analysis of urban heat exposure.
2. Leverages state-of-the-art machine learning techniques for predictive modelling of heat-health outcomes, advancing the field of environmental health research.
3. A cross-disciplinary approach enriches the interpretation of data, linking climate science with public health implications.
4. Risk of sampling bias due to secondary data utilisation, which may influence the representativeness of findings.
5. The spatial resolution of datasets, particularly those capturing microclimatic urban variations, may limit the granularity of exposure assessments, affecting the precision in capturing heat stress metrics.

## Budget

1. **Software Licenses**: The budget encompasses the licensing costs for specialized software used in data analysis and model building, including statistical software, machine learning libraries, GIS software, and data visualization tools.
2. **Cloud-Based Computing Resources**: Recognizing the need for high-performance computing, particularly for predictive modeling, we will allocate a significant portion of the budget to Google Cloud Computing services. This will support the computational demands of machine learning algorithms and large-scale data processing, ensuring efficiency and scalability.
3. **Hardware**: The project will also invest in acquiring suitable hardware or subscribing to additional cloud-based computational services to support data analysis.
4. **Publication Fees**: We anticipate expenses related to publishing our findings in open-access, peer-reviewed journals to ensure wide dissemination of our research.
5. **Training and Capacity Building**: The budget provides for ongoing training to keep the research team updated with the latest developments in data science and climate-health research.

The budget aligns with the funding limits of the National Institutes of Health (NIH) grant. Considering the project's scale and scope, we may explore additional funding sources, including grants, research partnerships, or institutional collaborations, to fully realize our research objectives.

## Supervisors

This research will be supported by an outstanding team of advisors, each bringing their vast knowledge and expertise in the intertwined disciplines of health and climate science.

**Professor Matthew Chersich**, based at Trinity College and the University of Witwatersrand, offers a wealth of experience in public health research that is invaluable to our study, particularly the health-related aspects. His career, spanning over two decades, has been focused on medical and public health research in Africa, particularly on maternal health and HIV, and recently on climate change and health. He has an extensive academic background in clinical medicine and public health, contributing to 14 WHO guidelines or monologues and serving as a contributing author to the Africa chapter of the 6th Intergovernmental Panel on Climate Change report. He has published more than 175 papers in peer-reviewed journals and has a significant H-Index of 48.

**Professor Akbar Waljee** of the University of Michigan brings crucial experience in statistical modelling and machine learning, essential for our data analysis and predictive modelling. Born in Kenya and educated in the United States, Prof. Waljee leads several key data and healthcare initiatives at the University of Michigan and the VA Ann Arbor Healthcare System. His work primarily involves utilizing machine learning and deep learning techniques to enhance healthcare access, quality, and efficiency, particularly in resource-constrained settings. His innovative work in decision support systems and personalized care is set to revolutionize patient care in gastroenterology and liver disorders in under-resourced settings globally.

**Dr. Christopher Jack** from the University of Cape Town strengthens the climate aspects of our study with his extensive knowledge in climate science, ensuring a well-rounded and sophisticated understanding of the climate-health nexus. With a background in computer science and ocean/atmospheric science, Dr. Jack possesses a unique blend of skills in high performance computing, modeling, analysis, science-society engagement, and decision-making under uncertainty. His current research activities are concentrated on the intersection of urban contexts and climate risk, leveraging his comprehensive expertise in climate science and modeling, and his proficiency in decision support and capacity development. His passion lies in working with and across diverse disciplines in complex problem spaces, making him especially interested in urban climate resilience in developing contexts.

Together, these advisors contribute a multidisciplinary perspective to our research, enriching its depth and breadth, and enhancing its potential impact.

## Conclusion

This research project endeavors to unravel the complex relationship between urban heat exposure, population vulnerability, and health outcomes within the unique socio-economic, environmental, and climatic context of Johannesburg. By employing advanced statistical techniques, machine learning methods, and a variety of robust data sources, the study aims to establish a nuanced understanding of the effects of heat on health in the city. This will culminate in the development of a spatially and demographically stratified heat-health outcome prediction model, which will enhance the city's readiness and response to heat-related health risks.

The insights generated from this study are expected to provide pivotal contributions to the fields of climate science, public health, artificial intelligence, and the interdisciplinary domain of climate and health. By advancing our knowledge and predictive capabilities, the research aims to inform targeted interventions and policies that reduce the adverse health impacts of heat exposure, particularly among vulnerable populations. As global temperatures continue to rise, the findings from this study could play a critical role in shaping public health strategies and urban planning efforts, not only in Johannesburg but also in other urban centers facing similar challenges.

Ultimately, this research seeks to contribute to the development of more resilient and healthy urban environments, where the risks associated with heat exposure are effectively managed and mitigated. By fostering a deeper understanding of the interplay between climate, health, and urban development, the project aims to pave the way for innovative solutions that ensure the well-being of communities in the face of escalating climate change

## References

Uncategorized References

1. Gasparrini, A., et al., *Mortality risk attributable to high and low ambient temperature: a multicountry observational study.* Lancet, 2015. **386**(9991): p. 369-75.

2. Analitis, A., et al., *Synergistic Effects of Ambient Temperature and Air Pollution on Health in Europe: Results from the PHASE Project.* Int J Environ Res Public Health, 2018. **15**(9).

3. Manyuchi A, et al., *Extreme heat events, high ambient temperatures and human morbidity and mortality in Africa: A systematic review.* Environmental Research (under review), 2020.

4. Chersich, M.F., et al., *Associations between high temperatures in pregnancy and risk of preterm birth, low birth weight, and stillbirths: systematic review and meta-analysis.* BMJ, 2020. **371**: p. m3811.

5. Rohat, G., et al., *Projections of Human Exposure to Dangerous Heat in African Cities Under Multiple Socioeconomic and Climate Scenarios.* Earth's Future, 2019. **7**(5): p. 528-546.

6. Watts, N., et al., *The 2020 report of The Lancet Countdown on health and climate change: responding to converging crises.* Lancet, 2020.

7. *IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)].*

8. Engelbrecht F, et al., *Projections of rapidly rising surface temperatures over Africa under low mitigation.* 10, 2015. **8**.

9.  *Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa). Retrieved from* [*https://commonfund.nih.gov/AfricaData*](https://commonfund.nih.gov/AfricaData)*.*

10. Azongo, D.K., et al., *A time series analysis of weather variability and all-cause mortality in the Kasena-Nankana Districts of Northern Ghana, 1995-2010.* Glob Health Action, 2012. **5**: p. 14-22.

11. Kovats RS, Wilkinson P, and Mohamed H, *Weather and cause-specific mortality in Cape Town, South Africa.* Epidemiology, 2005. **16**(5): p. S47-48. Program and Abstracts: The Seventeenth Conference of the International Society for Environmental Epidemiology (ISEE).

12. Kynast-Wolf, G., et al., *Seasonal patterns of cardiovascular disease mortality of adults in Burkina Faso, West Africa.* Trop Med Int Health, 2010. **15**(9): p. 1082-9.

13. Leone, M., et al., *A time series study on the effects of heat on mortality and evaluation of heterogeneity into European and Eastern-Southern Mediterranean cities: results of EU CIRCE project.* Environ Health, 2013. **12**: p. 55.

14. Scovronick, N., et al., *The association between ambient temperature and mortality in South Africa: A time-series analysis.* Environ Res, 2018. **161**: p. 229-235.

15. Wichmann, J., *Heat effects of ambient apparent temperature on all-cause mortality in Cape Town, Durban and Johannesburg, South Africa: 2006-2010.* Sci Total Environ, 2017. **587-588**: p. 266-272.

16. Asamoah, B., T. Kjellstrom, and P.O. Ostergren, *Is ambient heat exposure levels associated with miscarriage or stillbirths in hot regions? A cross-sectional study using survey data from the Ghana Maternal Health Survey 2007.* Int J Biometeorol, 2018. **62**(3): p. 319-330.

17. Bakhtsiyarava, M., K. Grace, and R.J. Nawrotzki, *Climate, Birth Weight, and Agricultural Livelihoods in Kenya and Mali.* Am J Public Health, 2018. **108**(S2): p. S144-s150.

18. MacVicar, S., et al., *Whether weather matters: Evidence of association between in utero meteorological exposures and foetal growth among Indigenous and non-Indigenous mothers in rural Uganda.* PLoS One, 2017. **12**(6): p. e0179010.

19. Diboulo, E., et al., *Weather and mortality: a 10 year retrospective analysis of the Nouna Health and Demographic Surveillance System, Burkina Faso.* Glob Health Action, 2012. **5**: p. 6-13.

20. Egondi, T., et al., *Time-series analysis of weather and mortality patterns in Nairobi's informal settlements.* Glob Health Action, 2012. **5**: p. 23-32.

21. Musengimana, G., et al., *Temperature Variability and Occurrence of Diarrhoea in Children under Five-Years-Old in Cape Town Metropolitan Sub-Districts.* Int J Environ Res Public Health, 2016. **13**(9).

22. Sewe, M., et al., *The association of weather variability and under five malaria mortality in KEMRI/CDC HDSS in Western Kenya 2003 to 2008: a time series analysis.* Int J Environ Res Public Health, 2015. **12**(2): p. 1983-97.

23. Johnson, D.P., J.S. Wilson, and G.C. Luber, *Socioeconomic indicators of heat-related health risk supplemented with remotely sensed data.* International Journal of Health Geographics, 2009. **8**(1): p. 57.

24. Jung, J., et al., *Heat illness data strengthens vulnerability maps.* BMC Public Health, 2021. **21**(1): p. 1999.

25. Xu, R., et al., *Socioeconomic level and associations between heat exposure and all-cause and cause-specific hospitalization in 1,814 Brazilian cities: A nationwide case-crossover study.* PLoS Medicine, 2020. **17**(10): p. e1003369.

26. Mhedhbi, Z., et al., *Mining the Web of Science for African cities and climate change (1991–2021).* Frontiers in Sustainable Cities, 2023. **5**.

27. Chersich, M.F., et al., *Impacts of Climate Change on Health and Wellbeing in South Africa.* Int J Environ Res Public Health, 2018. **15**(9).

28. Naicker, N., et al., *Indoor Temperatures in Low Cost Housing in Johannesburg, South Africa.* Int J Environ Res Public Health, 2017. **14**(11).

29. Ncongwane, K.P., et al., *A Literature Review of the Impacts of Heat Stress on Human Health across Africa.* Sustainability, 2021. **13**(9): p. 5312.

30. Pasquini, L., et al., *Emerging climate change-related public health challenges in Africa: A case study of the heat-health vulnerability of informal settlement residents in Dar es Salaam, Tanzania.* Sci Total Environ, 2020. **747**: p. 141355.

31. Wright, C.Y., et al., *Socio-economic, infrastructural and health-related risk factors associated with adverse heat-health effects reportedly experienced during hot weather in South Africa.* Pan Afr Med J, 2019. **34**: p. 40.

32. Enete, I., *Assessment of Health Related Impacts of Urban Heat Island (UHI) in Douala Metropolis, Cameroon.* International Journal of Environmental Protection and Policy, 2014. **2**: p. 35.

33. UNAIDS, *Seizing the moment. Global AIDS update. Tackling entrenched inequalities to end epidemics.* <https://www.unaids.org/en/resources/documents/2020/global-aids-report>, 2020.

34. UNAIDS, *UNAIDS data 2020.* <https://www.unaids.org/en/resources/documents/2020/unaids-data>, 2020.

35. Lubis, N., et al., *Prospective cohort study showing changes in the monthly incidence of Pneumocystis carinii pneumonia.* Postgrad Med J, 2003. **79**(929): p. 164-6.

36. *Worldometer. (2023, January 1). World population by country. Retrieved from* [*https://www.worldometers.info/world-population/population-by-country/*](https://www.worldometers.info/world-population/population-by-country/)*.*

37. Souverijns, N., et al., *Urban heat in Johannesburg and Ekurhuleni, South Africa: A meter-scale assessment and vulnerability analysis.* Urban Climate, 2022. **46**: p. 101331.

38. Tyson, P.D. and R.A. Preston-Whyte, *Weather and climate of southern Africa*. 2000.

39. Nana, M., K. Coetzer, and C. Vogel, *Facing the heat: initial probing of the City of Johannesburg’s heat-health planning.* South African Geographical Journal, 2019. **101**: p. 1-16.

40. Venter, Z.S., et al., *Green Apartheid: Urban green infrastructure remains unequally distributed across income and race geographies in South Africa.* Landscape and Urban Planning, 2020. **203**: p. 103889.

41. Chen, M., et al., *Rising vulnerability of compound risk inequality to ageing and extreme heatwave exposure in global cities.* npj Urban Sustainability, 2023. **3**(1): p. 38.

42. Khine, M.M. and U. Langkulsen, *The Implications of Climate Change on Health among Vulnerable Populations in South Africa: A Systematic Review.* Int J Environ Res Public Health, 2023. **20**(4).

43. Wright, C.Y., et al., *Major climate change-induced risks to human health in South Africa.* Environmental Research, 2021. **196**: p. 110973.

44. Gronlund, C.J., *Racial and socioeconomic disparities in heat-related health effects and their mechanisms: a review.* Curr Epidemiol Rep, 2014. **1**(3): p. 165-173.

45. *Centers for Disease Control and Prevention. (n.d.). South Africa. Retrieved August 3, 2023, from* [*https://www.cdc.gov/globalhealth/countries/southafrica/default.htm*](https://www.cdc.gov/globalhealth/countries/southafrica/default.htm)*.*

46. López-Carr, D., et al., *Extreme Heat and COVID-19: A Dual Burden for Farmworkers.* Front Public Health, 2022. **10**: p. 884152.

47. Alonso, L. and F. Renard, *A Comparative Study of the Physiological and Socio-Economic Vulnerabilities to Heat Waves of the Population of the Metropolis of Lyon (France) in a Climate Change Context.* International Journal of Environmental Research and Public Health, 2020. **17**(3): p. 1004.

48. *Gauteng City-Region Observatory (2019). Quality of life in the Gauteng city-region: A report on key indicators. Retrieved from* [*https://www.gcro.ac.za/about/annual-reports/*](https://www.gcro.ac.za/about/annual-reports/)*.*

49. National Institute of Statistics of Côte, d.I., *National Institute of Statistics of Côte d'Ivoire Datasets*. INS.

50. Hofierka, J., M. Gallay, and K. Onačillová, *Physically-based land surface temperature modeling in urban areas using a 3-D city model and multispectral satellite data.* urban climate, 2020. **31**: p. 100566.

51. Hooker, J., G. Duveiller, and A. Cescatti, *A global dataset of air temperature derived from satellite remote sensing and weather stations.* Scientific Data, 2018. **5**(1): p. 180246.

52. Kershaw, P., et al. *Delivering resilient access to global climate projections data for the Copernicus Climate Data Store using a distributed data infrastructure and hybrid cloud model*. 2019.

53. *Copernicus Climate Data Store (CDS)*. 2024, Copernicus Climate Change Service (C3S).

54. *Earth System Grid Federation (ESGF)*. 2024, ESGF.

55. Albrecht, C.M., et al., *Pairs (Re)Loaded: System Design & Benchmarking For Scalable Geospatial Applications.* 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), 2020: p. 488-493.

56. *10 m WorldCover 2020 v100*. 2021, European Space Agency (ESA).

57. *The Global Human Settlement Layer 2019 (GHSL 2019) public release*. 2021, Publications Office of the European Union, Luxembourg.

58. Arifwidodo, S.D., P. Ratanawichit, and O. Chandrasiri. *Understanding the Implications of Urban Heat Island Effects on Household Energy Consumption and Public Health in Southeast Asian Cities: Evidence from Thailand and Indonesia*. 2020.

59. Schwartz, R., et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* 2022.

60. Drobnič, F., A. Kos, and M. Pustišek, *On the Interpretability of Machine Learning Models and Experimental Feature Selection in Case of Multicollinear Data.* Electronics, 2020. **9**(5): p. 761.

61. Chen, R.-C., et al., *Selecting critical features for data classification based on machine learning methods.* Journal of Big Data, 2020. **7**(1): p. 52.

62. Uddin, S., et al., *Comparing different supervised machine learning algorithms for disease prediction.* BMC Medical Informatics and Decision Making, 2019. **19**(1): p. 281.

63. Sarica, A., A. Cerasa, and A. Quattrone, *Random Forest Algorithm for the Classification of Neuroimaging Data in Alzheimer's Disease: A Systematic Review.* Frontiers in Aging Neuroscience, 2017. **9**.

64. Xie, W., et al., *Predicting the Easiness and Complexity of English Health Materials for International Tertiary Students With Linguistically Enhanced Machine Learning Algorithms: Development and Validation Study.* JMIR Med Inform, 2021. **9**(10): p. e25110.