Cheating Charon: Predicting Excess Mortality Due to Temperature Variation Globally

Fernando Corral Lozada
f.corral@students.hertie-school.org

Paul Sharratt
p.sharratt@students.hertie-school.org

Rodrigo Filippi Dornelles
r.Dornelles@students.hertie-school.org

Abstract

Predicting the consequences of our changing climate for public health is a crucial area where machine learning can assist policy makers. As the effects of man-made climate change are felt, excess deaths due to extreme temperatures and weather events are expected to rise. However, due to the complexity of modelling the epidemiological relationship between populations and their habitats, evidence of this direct impact is limited. Climate change can directly and indirectly impact human health in multiple ways. Increasing food insecurity, the spread of disease vectors, armed conflicts, and extreme weather events can all be attributed to our changing climate. Aside from these indirect effects, the health consequences directly associated with variation in outdoor temperature are a key direct consequence. For the sake of feasibility and interpretability, we have chosen to focus on variations in air temperature as a key factor in climate change-related mortality in this paper. As machine learning methodologies mature, the scope of models' predictive capacities increases. For our project[1]*, we have chosen to test and compare predictive regression models to anticipate the number of excess deaths due to temperature variations caused by climate change. Our findings highlight the difficulties in generating reliable models and the range in efficacy, interpretability, and robustness of the models that can be applied to this problem. Moreover, we have applied our selected models to two separate, yet credible climate scenarios, further demonstrating the complexity of climate-change mitigation and adaptation for public health authorities and bodies. The public health consequences related to variation in air temperature have been explored in several studies and our project builds on this existing research, particularly the work of Honda et al (2014).[2]*

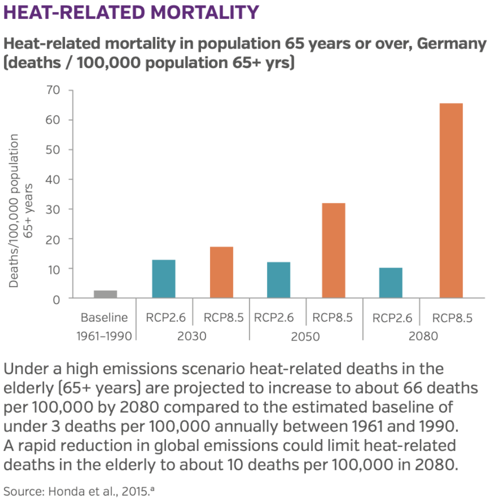
1. Introduction

Studying data science at a public policy school means looking for ways to use data to address the challenges that societies face. Amongst the myriad problems facing the world today, climate change is perhaps the most important and existential. It will impact all countries and have profound effects on the world's economies, societies, and basic habitability. The challenges posed by climate change are not abstract, but are something we all face in our day-to-day lives. The Hertie School, for instance, has sought to foreground questions of climate change through the Centre for Sustainability and the Sustainable Campus Initiative.

In 2002, the World Health Organization reported for the first time on the health impacts of climate change [3]. This report, however, did not include predictions of excess deaths due to heat-related impacts and temperature variation because of the complexity of modelling a relationship between ambient temperature and mortality. Machine learning approaches have matured significantly since the publication of that report and data on the impact of man-made climate change has become more readily and freely available to the research community. Machine learning is an important tool for citizens and policy makers to better understand the impact of climate change. Using machine learning, we can simulate scenarios, estimate the influence of different factors, predict certain consequences, and recommend mitigation measures.

We assume that deaths as a consequence of climate change, as illustrated in the figure above, are unnecessary, meaning that appropriate public policies might prevent them. This assumption is key to our motivation and our desire to produce the best results possible. As such, our project contributes to and itself demonstrates an increasing awareness of the significant and global consequences of inaction. We are motivated to study the implications of climate change, especially extreme temperature events, [6] and how they impact vulnerable populations. For example, in the summer of 2022, Berlin experienced an unprece-

[1] Link to our GitHub repo: https://github.com/rfdornelles/mds_ML_project



Under a high emissions scenario heat-related deaths in the elderly (65+ years) are projected to increase to about 66 deaths per 100,000 by 2080 compared to the estimated baseline of under 3 deaths per 100,000 annually between 1961 and 1990. A rapid reduction in global emissions could limit heat-related deaths in the elderly to about 10 deaths per 100,000 in 2080.

Source: Honda et al., 2015.[a]

Figure 1. Predicted heat-related Mortality in Germany[7]

dented heat wave that led to loss of life ([2]) Such extreme heats will become more commonplace without urgent action.

In our project, we applied different forecasting methods to our data and assessed their suitability to the problem of predicting the consequences of climate change. [1]. We fitted, trained, and tuned three separate models and have identified one in particular that presents the best results.

Finally, in terms of our own narrow motivation and from a pedagogical perspective, we believe that the work we have concluded on this project, the wealth of data we have explored, and existing methodologies that we have applied to our problem have enabled us to orient better ourselves at the intersection between climate change-related policy making and data science. As a result, we look forward to presenting our results to the Center for Sustainability and the Hertie School community at large, sparking debates about the importance of tackling this particular aspect of climate change with data-driven solutions.

2. Related Work

There is a considerable body of research that seeks to predict the consequences of climate change, be they direct or indirect, using machine learning. As such, identifying and understanding the research most relevant to problem was a considerable task in and of itself. Our work drew on two papers in particular, namely ”Fluctuating temperature modifies heat-mortality association around the globe” by Gasparrini et al and ”Heat-related mortality risk model for climate change impact projection” by Honda et al. Crucially, these two papers led us to narrow our focus on temperature variation and heat-related mortality, and provided us with a method for calculating temperature variation. Importantly, our decision to expand the scope of our project whilst simultaneously reducing its depth - by including a number of countries other than Germany into our model and opting to consider total population temperature-related mortality instead the heat-related mortality of particular age groups - was informed by the work of Gasparrini et al and Honda et al. Their work demonstrated the viability and efficacy of prediction models that incorporated comparatively few variables on a global scope.

Additionally, these papers demonstrated considerations that were outside the scope of our project work and what we could meaningfully achieve given time constraints and data limitations. In particular, we were unable to account for the localized V-shaped temperature–morality curve at national level, as explored in Honda et al[**?**], and the distributed lag effect of temperature changes over time, also explored in Honda et al. We recognise these two considerations as deficiencies in our model, but could not meaningfully account for them in this project. Ideally, this project will serve as the basis for future research, be it as a group or individually, and we hope that we will be able to address these limitations at a later date.

An important methodological reference point for the project was [1]. This paper in particular led us to our selection of models beyond the basic linear regression. This paper used four kinds of machine learning models including logistic regression, support vector machine, random forest, and XGBoost that were trained using variables selected for mortality prediction. Influenced by this approach, we decided to use Gradient Boost and XGBoost in our second round of models. However, as we did not have sufficient patient and public health data, it was not possible to go further in replicating this particular approach. Finally, through our research into the scientific literature on this topic, we recognised the imbalance of the data available, particularly climate and public health data, between industrialized or high-income countries and least developed countries.

3. Proposed Method

Our proposed method followed four overarching phases. These were:

1. Data Pre-processing
2. Data Cleaning & Structuring
3. Establishing Baseline Models
4. Tuning Selected Machine Learning Models

To achieve our goal of predicting how temperature impacts mortality, we identified several reliable public data sources and gathered a range of relevant data sets. By sourcing our data from recognised national and international institutions, such as the World Health Organization, World

[2] https://https://www.berlin.de/en/news/7634442-5559700-up-to-39-degrees-heat-wave-reaches-berli.en.html

Bank, and the Federal Statistical Office of Germany, we could be confident in the quality of our data. As students at a public policy school, our approach to this problem was informed by papers such as Simonsen et al (2021) [5] and Rai et al (2022) [4] that focus on the practical applicability of machine learning methodologies to policy problems, rather than a theoretical and academic approach. For the project as a whole, we collected data on:

- Historical and predicted population data for Germany
- Actual and predicted population in different age groups
- Mortality rate and main causes of death
- Temperature forecasting in line with the IPCC high-emissions scenario Representative Concentration Pathway (RCP) also called RCP8.5
- Temperature forecasting in line with low-emissions scenario, also called RCP4.5
- Emissions, specially CO2 and CFCs

However, it should be noted that we have not necessarily used all of the data that we collected. For the reasons explained below, we found that our work was ultimately productive when we were more selective with our data.

Temperature and meteorological data were obtained from Copernicus Climate Data Store. Mortality data and population projections from the World Bank and Our World in Data. We also obtained data sets on the causes of death amongst the different age groups globally. As detailed in further sections, we applied a series of regression equations to our data to quantify the current and historical relationships between mortality and a set of independent variables. As our data was sourced from several different sources and collections, we had to wrangle our data considerably.

The data we used for the final data set included:

- Causes of death from Our World in Data
- Global population from the World Bank
- Temperature data from the EU Copernicus Climate Data Store

After thoroughly researching the available data and its applicability to our problem, we created an initial version of the collated data frame that we could apply our models to. For this first data frame, we ran cross-validation on our regression models and came to the conclusion that our data set did not contain enough observations for a meaningful test/training split, as the data frame was limited to observations focused on public health and emissions metrics in Germany and between the years 2000 and 2019. The quantitative results for the R2 score varied considerably each time we ran our models, suggesting that the model was not robust enough. For this reason, we decided to increase our timescale by ten years, from 2000 to 1990, and to include environmental and public health data from several other countries. After expanding both our timescale and incorporating more countries into our data, we ran our baseline linear regression model again. Our results indicated that our data set was viable and that we could apply more models to it. Staying within regression, we opted to expand our model range to include other regression-based models. Finally, after testing these five models, we tuned the three best performing models and from these three identified the best overall performing model.

3.1. Data Cleaning

After our initial experiments with the second data set, we decided to restructure the data frame and focus on only what we believed to be the most important variables, as the range of environmental and emissions data (particularly the various greenhouse gasses) created considerable noise in our data set and increased our computation time. This final data set contained the following columns:

- The country
- The temperature
- Population
- Deaths due to climate conditions

The environmental data in this data set was sourced from the Copernicus Climate Data Store, which has the files in the .nc format, and map the temperature data to the coordinates of the countries in our sample. We then matched each country of interest with its coordinates in order to extract a mean annual temperature for each country. This was a complicated and time consuming process. Having achieved this, it was possible to merge the countries with the Word Bank Indicators data on population, matching each country with a specific year. We then merged the data set with the database from Our World in Data, containing the number of deaths of each main cause. This data enable us to isolate global exposure deaths due to heat and/or cold in our final data frame. It is important to acknowledge that, in the first version, we tried also to bring together data about emissions, age groups and mortality in Germany. This, however, led to many difficulties to find relevant features and made our data set excessively small. Despite the excessive time spent on data collection and wrangling, we are convinced that including more data from different countries will make our models perform better and less like to generate over-fitted results.

3.2. Data Preprocessing

The most important step for our data processing was building a pipeline. For this we used Scikit-Learn's open source machine learning library and applied a StandardScaler. We did this because our columns were in different units of measure, which - without a StandardScaler - could have generated confusion and distortions in our data. For our future work, we will try to figure out if strategies such as logarithmic transformation or polynomial transformation might help us achieve better results. This might be especially important if we decide to add more features, such as categorical values (to test different effects over different population groups).

3.3. Establishing Baseline Machine Learning Models

Since our main goal at this stage was to build a functional pipeline, we decided to test our data running basic linear models, in particular linear regression. Having proven that our data frame was robust enough to run a first baseline model, we supplemented our approach with further research in the scientific literature in order to apply more advanced methodologies to our project.

For this reason, we preferred to utilize references that were more applicable in a public policy context - rather than a medical or purely public health one. For instance, [2] presented an effective methodology to predict deaths using temperature as a dependent variable, applying a distributed Lag model.

In sum, we used the following five models on our data sets:

- Linear Regression
- Decision Tree
- Gaussian Process Regression
- Gradient Boosting Regressor
- Extreme Gradient Boosting Regressor

As we are ultimately investigating the effects of two separate climate scenarios (RCP 4.5 & RCP 8.5) on mortality using five different models, one key successful outcome of the project will be ten comparable predictions. The models were trained using the mortality rate data, demographic data, and the mean temperature. The first key evaluation metric for all five selected models was the R2 score. We chose the R2 score as we wanted to measure the proportion of the variance in the dependent variable (deaths due to temperature variations) that is predictable from the independent variables (public health data, emissions data, and environmental data) and assumed this would be a good metric for comparisons between models.

As we anticipated, the pipeline requirements varied somewhat from model to model, increasing the overall time we spent on this phase of the workflow. However, as our pedagogical goal is to learn as much as possible about the differences between the various models, we believe that this effort to be worthwhile.

After the data wrangling process, we separated all our data into training and testing data, using a 80% and 20% split, leaving aside the testing data and working solely with the training data. This final split was used across all of the models. We also carried out some feature engineering, such as scaling the data to normalize the range of independent variables and performing the necessary transformations such as also adding a random state value to the function.

Since our focus was in setting up working models, we were initially more concerned with good accuracy and avoiding the effects of over-fitting. The accuracy of the models seemed to be a good start, and our initial quantitative results were better than we expected, especially after changing the strategy and building a larger data set.

3.4. Establishing Baseline Models

Following the results of our initial round of models, we selected only the three best performing models, where were:

- Decision Tree
- Gradient Boosting Regressor
- Extreme Gradient Boosting Regressor

After data wrangling and running baseline models, the third main phase in our project was tuning the hyperparameters of the three best performing models from the second phase. We tuned the parameters for the Gradient Boost and XGBoost models referencing their particular guidelines to achieve the highest accuracy and precision scores. The parameters used for the Gradient Boosting Regressor were the number of estimators, maximum depth, minimum samples split, learning rate, and loss. For the Extreme Gradient Boosting Regressor, we imported RandomizedSearchCV into the script. We applied the following parameters: xgb model, colsample bytree, gamma, learning rate, max depth, n estimators, and subsample, setting a random state of 42 and 200 parameter setting tries (n iter), three cross validations, verbose, n jobs and return train score. Once the parameter tuning testing was done, the remaining testing data was introduced to the model to predict the mortality rate and temperature.

4. Experiments

Data: We sourced data sets from the Federal Statistical Office of Germany, World Health Organization, the European mortality database, and the European Commission's Eurostat for this project.[7] As our data was sourced from several different sources and collections, our data required considerable wrangling before we could created our training and test sets and apply our models.

Model	R2	AdjR2	RMSE	MSE
Linear Regression	0.18	0.17	600	360686
Decision Tree	0.95	0.95	145	21031
Gaussian Process	0.65	0.65	389	151988
Gradient Boost	0.97	0.97	110	12115
XGBoost	0.96	0.96	126	16107

Table 1. Models & Evaluation Scores

Software: For both hosting our code and coordinating our project work, we used GitHub, Google Collab, and Teams, and we worked using Python (to do the modeling and visualization) and also used R (only to support the wrangling tasks). Beyond access to Google Collab servers, our hardware requirements were minimal.

Evaluation method: Since we applied regression models - and following our instructor's feedback - we were concerned about applying evaluation methods that could be comparable. We used the R2 score as our initial evaluation metric. In the final output, we used the following metrics:

- R2 Score
- Adjusted R2 Score
- Root-Mean-Square Error (RMSE)
- Mean Square Error

We chose the R2 score as a basic measure of fit that could easily be derived for each model. In the first analysis and when devising our approach, this was the first and only evaluation metric we considered. We did not incorporate the remaining four evaluation metrics until we had a working pipeline for the four chosen models. Happily though, the metrics we identified apply across a range of regression and prediction models, so we were not as limited in our choice as we had feared we might be. The Adjusted R2 score was a natural choice, as it compensates for the addition of variables and only increases if the new term enhances the model above what would be obtained by probability and decreases when a predictor enhances the model less than what is predicted by chance. This was an important consideration as it enabled us to compare our R-squared values and rule out over-fitted models. Finally, Mean Square Error and Root-Mean-Square Error (RMSE) In the final analysis, our best performing model was our Extreme Gradient Boosting Regressor model. However, our initial, untuned Gradient Boost model also performed well, on the basis of our chosen evaluation metrics. The Linear Regression and Gaussian Process models did not perform well, whilst the Decision Tree model initially performed well, but did not withstand our efforts to tune it. This was one of the most surprising results and probably one key educational takeaway from the project as whole, namely the dramatic change in results when hyperparameter tuning does not have its desired effect.

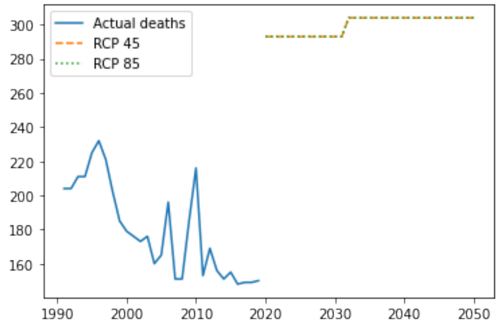


Figure 2. Results from Extreme Gradient Boosting Regressor Model using actual deaths and predicted deaths

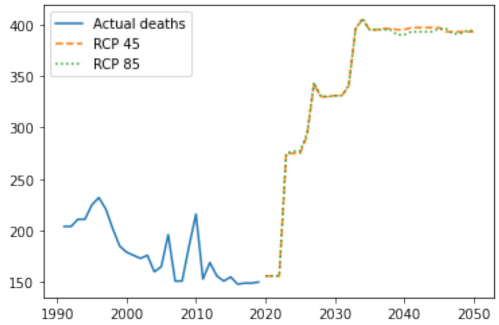


Figure 3. Results from the Gradient Boost Model displaying actual deaths and predicted deaths

Results: Our results for the evaluation of the various models can be seen in Tables 1 and 2, whilst the results for the predictions of our best performing models, the Gradient Boost and XGBoost, can be seen in Figures 2 and 3. Unfortunately, the predictive results from the models are not totally reliable, since they suggest a excessively-quick raise in deaths and, in both cases, doesn't suggest taking into consideration the differences between both scenarios.

Our dependent variable throughout the models was the quantity of deaths caused by exposure. Our independent variables were temperature and population.

Our quantitative results for the initial selection of models

Tuned Model	R2	AdjR2	RMSE	MSE
Decision Tree	0.00	0.00	664	441096
Gradient Boost	0.96	0.96	118	13944
XGBoost	0.96	0.96	126	16107

Table 2. Tuned Models & Evaluation Scores

was somewhat surprising. The difference between the three best performing models and the linear regression and Gaussian process models was quite pronounced. However, we had intuited from our class work that Decision Tree, Gradient Boost, XGBoost might be well suited to this particular prediction task. The result of the linear regression model was probably to be expected, but was not of especial significance as we had primarily used the model to establish the validity of our data structure, test/train splits, and a baseline for comparison with the other models. In that sense, it was quite beneficial that the linear regression model perfomed as poorly as it did. Understandably, we had hoped that our tuning efforts would improve the performance of our three chosen models. For that reason, we were very surprised by the results of our tuned Decision Tree model. Clearly, our tuning did not have the desired effect, and we are frankly still unclear as to the exact issue. Similarly, after the tuning our Gradient Boost suffered a slight decrease in performance, whereas the XGBoost remained the same. Nevertheless, the two gradient models performed well quantitatively.

5. Analysis

On their most basic level, our models demonstrate a relationship between deaths and temperature. However, our models are unfortunately not precise enough to indicate when a rise in deaths will occur. Whilst the models correctly predict a positive relationship between deaths and temperature, they aren't sensitive or accurate enough to demonstrate this relationship under the circumstances of differing climate pathways. Clearly, for any future attempt at addressing this problem, we would consider incorporating some time series element into models and approach and expanding our data sets to incorporate more variables, particularly emissions variables and other environmental data, unique to each particular climate pathway.

During our first attempt to develop a model, where we focused purely on mortality predictions for Germany, the models proved to be quite aggressive and, applying some degree of domain knowledge, not very accurate. Expanding the scope of all our models and tuning the hyperparameters on our three best performing models made our results closer in line with reasonable domain expectations for future climate-related mortality and with other research. However, owing to the constraints of the project, we could not tune our models further, nor could we fully identify the most relevant features. This is also a key insight into the performance of our models. We began this project with considerable environmental and emissions data specific to Germany. We could not develop a technique to identify the best predictors, so we made a judgement call as to what we believed or expected would be the most important variable, namely air temperature. This trade off was difficult to make. However, owing in part to our ambitions for our models and, more importantly, to the availability of commensurate climate and emissions data for countries other than Germany, we believe that it was the right choice both to focus on temperature as our predictor and consequently to expand our data to incorporate data from a wider range of countries.

Assuming that air temperature is a key predictor of excess mortality is not unreasonable, particularly given the existing literature that confirms our hypothesis. However, this assumption does necessarily simplify a complex picture, one that was already complicated by our desire to predict outcomes under two distinct, but closely related scenarios. As the closeness of the lines in the plots above indicate, our models were not able to capture the difference in mortality outcomes between the two scenarios effectively. Again, applying a degree of domain knowledge, it is not difficult to arrive at the conclusion that this is a failing of our predictions.

6. Conclusions

Overall, we are satisfied that we were able to test five separate models on our data, realize the limitations of the models through meaningful evaluation metrics, subset and at least partly tune the three best models, and finally identify and further tune the most effective model within our initial selection. In the context of our chosen policy domain, we believe that we have effectively illustrated the varying effectiveness of widely-used machine learning models in answering our guiding question. In this sense, our project illustrates the difficulties of applying machine learning models to an extremely complex task. We aware that we have simplified the problem for the sake of our project. In terms of our pedagogical aims, however, we have all learnt considerably from the project, be it from the necessary, but exhaustive data wrangling process to the fine-tuning of models through to hyperparameter optimization and communication of our results.

Whilst we are of the opinion that our project was a qualified success, we would like to highlight a few areas where we faced difficulty and some weaknesses in our final result. We hope these limitations and deficiencies could serve as the basis for future work. First, foremost and perhaps most predictably, we believe our data could incorporate more variables, particularly from public health and climate data sets. In general, we underestimated the challenge of finding,

Year	Actual	M1_rcp45	M1_rcp85	M2_rcp45	M2_rcp85
2015	155.0	-	-	-	-
2016	148.0	-	-	-	-
2017	149.0	-	-	-	-
2018	149.0	-	-	-	-
2019	150.0	-	-	-	-
2020	-	293.0	293.0	156.0	156.0
2021	-	293.0	293.0	156.0	156.0
2022	-	293.0	293.0	156.0	156.0
2023	-	293.0	293.0	275.0	275.0
2024	-	293.0	293.0	275.0	277.0

Table 3. Number of deaths in Germany due heat or cold: sample of outcomes from M1 (XGBoost Model) and M2 (Gradient Boost Model) in different scenarios, comparing also the actual values (until 2019)

wrangling, and structuring appropriate data for this project. In particular, we struggled to find usable climate data on the different climate pathways. The problem was not so much the availability of the data, but rather its volume and complexity. The data sets we were able to find included a large number of emissions variables. Temperature variation or average temperatures are a gross over-simplication of the complexity of climate data and predicted scenarios. The dependent variables of climate change for mortality should not in our view be reduced solely to temperature variation. Reducing the problem to only this aspect fails to reflect the enormity of the problem and its risks. However, a more nuanced picture was beyond the scope of this project and we are satisfied that we have both proven the difficultly in generating robust models and demonstrated the importance and utility of predicting excess deaths for different scenarios in climate-change policy-making.

7. Contributions

In this project, we discussed and planned our approach together before assigning work to individuals. We used an Agile model using a Kanban board and weekly meetings to manage and assign tasks. Using this flexible project management approach, we were able to focus individually on different areas according to our respective skill sets and to the project's workflow.

At the outset, the team members contributed equally to the definition of our problem and our goals. Additionally, all team members researched relevant papers and sourced the data collections and sets that can be found in our GitHub repository. We divided the remaining tasks as follows: Fernando Corral Lozada, as our most experienced Python coder, assumed overall responsibility for our models; Rodrigo Dornelles, who has considerable prior experience with data wrangling in R, was primarily responsible for wrangling and structuring our data; whilst Paul Sharratt assisted with data wrangling, model evaluation and tuning, and with the writing of the reports.

Each team member took responsibility for data collection and wrangling on for specific domain areas. Paul was responsible for the public health data sets that we used, whilst Fernando managed and sourced global and national climate data. Finally, Rodrigo managed and sourced data on relevant demographics and populations.

References

[1] Y. Hirano, Y. Kondo, T. Hifumi, S. Yokobori, J. Kanda, J. Shimazaki, K. Hayashida, T. Moriya, M. Yagi, S. Takauji, J. Yamaguchi, Y. Okada, Y. Okano, H. Kaneko, T. Kobayashi, M. Fujita, H. Yokota, K. Okamoto, H. Tanaka, and A. Yaguchi. Machine learning-based mortality prediction model for heat-related illness. *Scientific Reports*, 11(1):9501, Dec. 2021.

[2] Y. Honda, M. Kondo, G. McGregor, H. Kim, Y.-L. Guo, Y. Hijioka, M. Yoshikawa, K. Oka, S. Takano, S. Hales, et al. Heat-related mortality risk model for climate change impact projection. *Environmental health and preventive medicine*, 19(1):56–63, 2014.

[3] W. H. Organization. *The world health report 2002: reducing risks, promoting healthy life*. World Health Organization, 2002.

[4] M. Rai, S. Breitner, K. Wolf, A. Peters, A. Schneider, and K. Chen. Future temperature-related mortality considering physiological and socioeconomic adaptation: a modelling framework. *The Lancet Planetary Health*, 6(10):e784–e792, 2022.

[5] L. Simonsen and C. Viboud. Mortality: A comprehensive look at the covid-19 pandemic death toll. *Elife*, 10:e71974, 2021.

[6] C. Winklmayr, S. Muthers, H. Niemann, H.-G. Mücke, and M. an der Heiden. Heat-related mortality in Germany from 1992 to 2021. *Deutsches Ärzteblatt international*, July 2022.

[7] World Health Organization and United Nations Framework Convention on Climate Change. Health and climate change: country profile 2021: Iraq. Technical report, World Health Organization, Geneva, 2021. Section: 15 p. WHO/HEP/ECH/CCH/21.01.10.