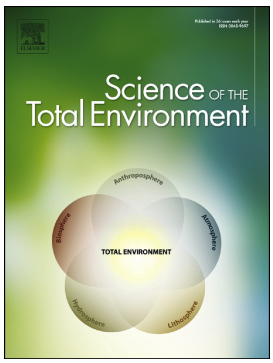
Journal Pre-proof



Machine and deep learning for modelling heat-health relationships

Jeremie Boudreault, Celine Campagna, Fateh Chebana

PII: S0048-9697(23)03283-7

DOI: https://doi.org/10.1016/j.scitotenv.2023.164660

Reference: STOTEN 164660

To appear in: *Science of the Total Environment*

Received date: 2 May 2023

Revised date: 22 May 2023

Accepted date: 2 June 2023

Please cite this article as: J. Boudreault, C. Campagna and F. Chebana, Machine and deep learning for modelling heat-health relationships, *Science of the Total Environment* (2023), https://doi.org/10.1016/j.scitotenv.2023.164660

Machine and deep learning for modelling heat-health relationships

Jeremie Boudreault[ab*], Celine Campagna[ab] and Fateh Chebana[a]

[a]Centre Eau Terre Environnement, Institut national de la recherche scientifique (INRS), 490 de la Couronne, Québec (QC) Canada, G1K 9A9

[b]Direction de la santé environnementale, au travail et de la toxicologie, Institut national de santé publique du Québec (INSPQ), 945 avenue Wolfe, Québec (QC) Canada, G1V 5B3

* Corresponding author :
Jeremie Boudreault (jeremie.boudreault@inrs.ca)
490 de la Couronne, Québec (QC) Canada, G1K 9A9

Revised manuscript submitted to Science of the Total Environment

May 22, 2023

Abstract

Extreme heat events pose a significant threat to population health that is amplified by climate change. Traditionally, statistical models have been used to model heat-health relationships, but they do not consider potential interactions between temperature-related and air pollution predictors. Artificial intelligence (AI) methods, which have gained popularity for health applications in recent years, can account for these complex and non-linear interactions, but have been underutilized in modelling heat-related health impacts. In this paper, six machine and deep learning models were considered to model the heat-mortality relationship in Montreal (Canada) and compared to three statistical models commonly used in the field. Decision Tree (DT), Random Forest (RF), Gradient Boosting Machines (GBM), Single- and Multi-Layer Perceptron (SLP and MLP), Long Short-Term Memory (LSTM), Generalized Linear and Additive models (GLM and GAM), and Distributed Lag Non-Linear Model (DLNM) were employed. Heat exposure was characterized by air temperature, relative humidity and wind speed, while air pollution was also included in the models using five pollutants. The results confirmed that air temperature at lags of up to 3 days was the most important variable for the heat-mortality relationship in all models. NO_2 concentration and relative humidity (at lags 1 to 3 days) were also particularly important. Ensemble tree-based methods (GBM and RF) outperformed other approaches to model daily mortality during summer months based on three performance criteria. However, a partial validation during two recent major heatwaves highlighted that non-linear statistical models (GAM and DLNM) and simpler decision tree may more closely reproduce the spike of mortality observed during such events. Hence, both machine learning and statistical models are relevant for modelling

heat-health relationships depending on the end user goal. Such extensive comparative analysis should be extended to other health outcomes and regions.

Keywords : mortality, temperature, humidity, air pollution, machine learning, deep learning.

1. Introduction

In its latest report (IPCC, 2021), the Intergovernmental Panel on Climate Change (IPCC) reaffirmed that climate change increases the frequency, intensity, length and spatial extent of many weather events such as extreme heat (Casati et al., 2013; Jeong et al., 2016; Meehl & Tebaldi, 2004). Extreme heat events pose a significant threat to population health because of their impact on both mortality (Basu, 2009; Basu & Samet, 2002; Gosling et al., 2009; Xu et al., 2016) and morbidity (Li et al., 2015; Ye et al., 2012), as well as important economic consequences (Wondmagegn et al., 2019).

The heat-health relationship is commonly studied using an over-dispersed Poisson regression statistical model (Gosling et al., 2009). Non-linear approaches using either splines (e.g., Doyon et al., 2008; Ishigami et al., 2008) or Generalized Additive Models (GAM) (e.g., Bayentin et al., 2010; S. Lin et al., 2012; Pascal et al., 2013) are usually preferred to linear ones i.e., Generalized Linear Models (GLM) (e.g., Basu et al., 2012; Schwartz et al., 2004). To describe both the lag structure and the non-linear effect of the exposure, the Distributed Lag Non-Linear Model (DLNM) was proposed (Armstrong, 2006; Gasparrini et al., 2010) and became very popular in the last decade (e.g., Gasparrini et al., 2015, 2017; Pascal et al., 2018, 2021; Vicedo-Cabrera et al., 2018, 2021).

The exposure to heat is characterized by a temperature variable at various lags, mainly the daily mean temperature (Son et al., 2019). Minimum or maximum temperature, as well as composite temperature indices (e.g., Humidex, Heat Index, Apparent Temperature, Wet Bulb Globe Temperature, etc.) can also be used (Barnett et al., 2010; Kovats & Hajat, 2008; Tong & Kan, 2011; Vaneckova et al., 2011; Zhang et al., 2014). It is still unclear if air pollution changes the heat-health relationship or not (Huang et al., 2011; Son et al., 2019). While some heat-health studies do not include air pollution (e.g., Barnett et al., 2010; Doyon et al., 2008), others include it only in their sensitivity analyses (e.g., Gasparrini et al., 2015). Because their levels are generally elevated during extreme heat events, air pollutants should be considered when studying the health effects of heat (Huang et al., 2011; Kovats & Hajat, 2008).

In the studies cited above, the temperature-related predictors and, when considered, air pollution variables, were always treated separately. Indeed, no interactions were included between these variables because of the difficulty of easily considering them in statistical models traditionally used. In contrast, machine and deep learning models, a branch of artificial intelligence (AI), can easily take these interactions into account. However, these models were only seldom used for modelling heat-health relationships (e.g., Y.-C. Lin et al., 2021; Masselot et al., 2021; Nishimura et al., 2021; Ogata et al., 2021; J. Park & Kim, 2018; M. Park et al., 2020; Y. Wang et al., 2019). In most applications, a single model was considered e.g., Random Forest (Y. Wang et al., 2019; Zhang et al., 2014) or Multi-Layer Perceptron (Khatri & Tamil, 2017). Only a few studies have so far compared more than two approaches (Marien et al., 2022; Nishimura et al., 2021; Ogata et al., 2021; M. Park et al., 2020; Qiu et al., 2020). No studies have yet compared the results of machine

and deep learning models with the widely used DLNM. Finally, recurrent neural networks such as the Long Short-Term Memory (LSTM) have only been seldom used (Y.-C. Lin et al., 2021; Nishimura et al., 2021).

In addition, the calibration of machine and deep learning models in above studies can be questioned. Models have been mostly calibrated on small datasets of <5 years (e.g., Khatri & Tamil, 2017; Ogata et al., 2021; Park et al., 2020; Qiu et al., 2020; Wang et al., 2019) or 5–10 years of data (e.g., Kassomenos et al., 2011; Lin et al., 2021; Nishimura et al., 2021; Zhang et al., 2014). Furthermore, a recent review of machine learning in public health reported that most studies failed to report their hyperparameters (Morgenstern et al., 2020). In some studies cited above, no hyperparameters optimization is performed at all (e.g., Zhang et al., 2014). Hence, subjective or unjustified choices of those hyperparameters can be expected, leading to a suboptimal fit of these models. Finally, air pollution was absent from most heat-health studies using machine and deep learning that primarily focused on temperature-related variables (e.g., Mora et al., 2017; Nishimura et al., 2021; Ogata et al., 2021; Park et al., 2020; Y. Wang et al., 2019; Zhang et al., 2014).

By allowing potential complex interactions between temperature-related variables and air pollution, machine and deep learning can lead to better performance for modelling the heat-health relationship. In this study, six machine and deep learning models such as tree-based methods, feedforward and recurrent neural networks were compared to statistical models commonly used in the field. A transparent calibration procedure (i.e., hyperparameters optimization) with a long-enough training dataset and an adequate selection of heat-related predictors including air pollution was used for the fitting of the models.

The paper is organized as follows. Section 2 contains the material and methods. The results are presented in section 3. Section 4 discusses the obtained results while Section 5 concludes this paper.

2. Material and Methods

This project received ethics approval from the Human Research Ethics Committee of the National Institute of Scientific Research (CER-22-693).

2.1. Data sources

The case study for analyzing the heat-health relationship with machine and deep learning focussed on the Census Metropolitan Area (CMA) of Montreal (Figure 1). The CMA of Montreal had a population of around 4.3 million inhabitants in 2021, which corresponds to approximately half the population of the province of Quebec, Canada (Statistics Canada, 2022).

The studied health outcome was the all-cause mortality. Mortality data from 1981 to 2019 (prior to the COVID-19 pandemic) in the CMA of Montreal was provided by the *Institut national de santé publique du Québec* (INSPQ). Seasonal and long-term trends in mortality time series were removed prior to modelling using a natural cubic spline of day of the year with 5 degrees of freedom for the seasonal trend, a linear function of year for the long-term trend and indicators for weekdays and holidays (Zhang et al., 2012; 2014). Other mortality trends were also tested (e.g., Gasparrini et al., 2010, 2015), but were not selected for further analyses as they did not differ significantly. The adjusted seasonal and long-term trends were then subtracted from the crude daily mortality to obtain the

response variable of interest for the models, namely the daily **mortality deviation**, i.e. the over- and under-mortality relative to the expected seasonal and long-term value (Zhang et al., 2012, 2014).

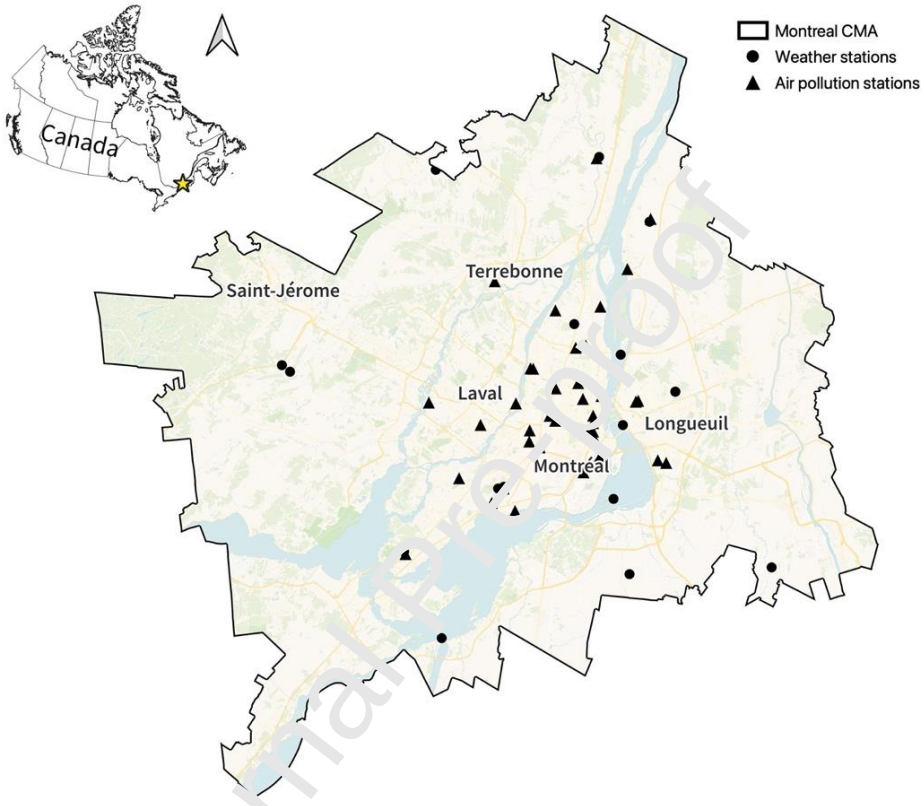


Figure 1: Weather and air pollution stations within the Census Metropolitan Area (CMA) of Montreal.

Weather and air pollution data were provided by Environment and Climate Change Canada (ECCC) and the National Air Pollution Surveillance (NAPS) program, respectively. Hourly data from weather and air pollution stations within Montreal CMA (Figure 1) were aggregated to mean daily values for the variables of interest. Temperature-related variables included air temperature, relative humidity and wind speed. These three variables are the ones commonly used to define composite

temperature indices such as humidex or apparent temperature (e.g., Barnett et al., 2010; Vaneckova et al., 2011). For air pollution, the two most studied variables in heat-mortality relationships were used (Son et al., 2019), namely ozone (O_3) and particulate matter of 2.5 microns or less in diameter ($PM_{2.5}$) as PM_{10} was not available. Other air pollutants more broadly used in weather-health studies such as nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and carbon monoxide (CO) were also considered (e.g., Basu et al., 2012; Goldberg et al., 2011; Lavigne et al., 2014; X. Wang et al., 2014).

Daily temperature-related and air pollution data at various stations were combined using a spatial mean to obtain a single time series for each variable for the entire studied region as done previously (e.g., Chiu et al., 2021; Masselot et al., 2018). Because the lagged effect of temperature/air pollution variables on health is important (Son et al., 2019), lags up to 7 days for all predictors were considered, with the following aggregates: value at lag 0, mean of values at lags 1 to 3 days and mean of values at lags 4 to 7 days. The period May 1st to September 30th was studied for the heat-mortality relationship. The overlapping period between mortality, weather and air pollution data was 1998–2019 (Table 1). This led to 3366 observations for that 22-year period. As $PM_{2.5}$ was the only variable missing for the 1981–1997 period (for which all other variables were available), another dataset, referred to as "supplementary dataset", consisting of years 1981–2019, but without $PM_{2.5}$ variable, was also created (Table S1). This supplementary dataset had 39 years of data for a total of 5967 observations.

Table 1: Overview of the data for modelling the heat-mortality relationship in Montreal CMA (1998–2019, May–September). All data are daily values.

	Min	Q10	Q25	Median	Mean	Q75	Q90	Max	Source
Mortality variables									
Crude mortality	37.00	56.00	61.00	68.00	68.37	75.00	81.00	135.00	INSPQ
Mortality deviation	-29.32	-12.07	-7.01	-1.06	-0.77	5.23	10.75	65.66	INSPQ
Predictors variables									
Air temperature (ºC)	3.21	11.79	15.24	18.64	18.13	21.21	23.54	28.97	ECCC
Relative humidity (%)	30.88	54.78	63.44	71.16	70.45	78.41	85.64	97.35	ECCC
Wind speed (km/h)	3.23	6.45	8.01	10.33	10.79	12.97	15.76	29.70	ECCC
O_3 concentration (ppb)	3.77	13.97	18.11	23.76	24.53	29.52	35.48	68.58	NAPS
NO_2 concentration (ppb)	1.37	4.84	6.42	9.05	10.04	12.64	16.76	39.52	NAPS
SO_2 concentration (ppb)	0.00	0.26	0.57	1.23	1.68	2.33	3.69	9.87	NAPS
CO concentration (ppm)	0.09	0.17	0.20	0.26	0.28	0.33	0.41	0.83	NAPS
$PM_{2.5}$ concentration ($\mu g/m^3$)	0.33	3.47	5.15	7.66	9.08	11.51	16.25	59.08	NAPS

The datasets were split into two distinct periods following the hold-out method for time series using 70% of the first years of data for the calibration (training) of the models and the remaining 30% for the validation (test) of the models. The training datasets had respectively the years 1998–2013 and 1981–2009 for the main (with all predictors) and the supplementary (without $PM_{2.5}$ predictor) datasets. The validation sets consisted of years 2014–2019 and 2009–2019 for the main and supplementary datasets, respectively.

2.2. Tree-based methods

Tree-based methods, also called classification and regression trees, come from both statistical and machine learning fields (James et al., 2013). A single tree is called a

Decision Tree (DT) because of its reversed tree shape (Quinlan, 1986). DT is easy to interpret and explain, but often lack prediction accuracy (James et al., 2013). Hence, ensemble methods that consist of several trees have been proposed. **Random Forest** (RF) is a model in which a forest of fully grown trees is built using bootstrapped datasets of observations as well as a subset of predictors for each node in the underlying DTs (Breiman, 2001). **Gradient Boosting Machines** (GBM) are also based on DT but differ from RF in their construction (Friedman, 2001). Trees with fewer leaves (e.g., 1 to 5) called weak learners are grown using a sequential fitting method instead of bigger trees built in parallel in RF. The next tree in GBM is fitted on the residual of the last tree(s) given a shrinkage parameter λ also called learning rate. As in RF, GBM also uses bootstrapped datasets as well as a subset of the available predictors for the underlying trees.

DT, RF and GBM were fitted to model daily mortality deviation as a function of (lagged) temperature-related and air pollution variables. In DT, a fully grown tree was first fitted. Then, a pruning procedure was applied to decrease the generalization error (James et al., 2013). The optimal amount of pruning (i.e., the number of leaves) was found using a 5-fold cross-validation on the training set. Because we had no past evidence on the optimal tuning of RF and GBM in the context of heat-health relationships, an extensive grid search procedure was performed. This method allowed us to test various combinations of hyperparameters and draw conclusions for further analysis, even though it can be less efficient than other hyperparameter optimalization methods such as random sampling (Bergstra & Bengio, 2012). DT, RF and GBM were fitted using *tree*, *randomForest* (Liaw & Wiener, 2002) and *gbm* (Greenwell et al., 2019) packages in R.

The unique hyperparameter for DT was tree depth. The hyperparameters for RF included the number of trees (500, 1000, 2500 and 5000) and the fraction of predictors considered at each tree split (square root of the number of predictors, 1/4, 1/3, 1/2, 3/4, and all predictors). The hyperparameters for GBM were learning rate (0.001, 0.01, and 0.1), tree depth (1, 3 and 5), number of trees (1000, 2500 and 5000) and the fraction of predictors at each split (1/3, 1/2 and 3/4). The hyperparameters grid search was performed using a 5-fold cross-validation on the training dataset to minimize out-of-sample root mean square error (RMSE) (Chapter 5 in Goodfellow et al., 2016).

To explain RF and GBM models, which are built using a large amount of DT (up to 5000 in our case), feature importance (FI) metrics were computed. FI is a standard method in machine learning to know which variables contribute the more to the prediction success (Chapters 15 and 16 in Hastie et al., 2009). For RF, two FI metrics were computed using 1) node purity, based on the mean decrease in mean square error (MSE) for each predictor in all underlying trees, and 2) permutation, based on the increase in out-of-bag prediction error when one predictor is randomly shuffled. For GBM, FI was computed from the decrease in MSE for each predictor (i.e., node purity). All FI metrics were scaled to a maximum value of 1 to allow comparison between different models/metrics.

2.3. Neural networks

Neural networks are machine and deep learning methods inspired by the human brain. One of the simplest neural networks is the feedforward **Single-Layer Perceptron** (SLP). It contains three layers: an input one, a hidden one and an output one. SLP is generalized into the **Multi-Layer Perceptron** (MLP) (Chapter 6 in Goodfellow et al., 2016). MLP

allows for more than one hidden layer and belongs, in this context, to the family of deep learning models. SLP and MLP were kept separated in this study for comparison purposes. While feedforward neural networks consider observations independently, recurrent neural networks, such as the **Long Short-Term Memory** (LSTM), transfer information from one cell to the other and are particularly adapted for sequential data such as time series (Chapter 10 in Goodfellow et al., 2016). LSTM extends classical recurrent networks by including a memory function and correcting the vanishing gradient problem (Hochreiter & Schmidhuber, 1997).

As for RF and GBM, no past evidence was found about the optimal tuning of neural networks in the specific context of heat-health relationships. Hence, an extensive grid search was performed. For SLP and MLP, hyperparameters included learning rate (0.0001, 0.001 and 0.01), activation function (Rectified Linear activation Unit (ReLU), hyperbolic tangent (tanh) and logistic) and scaling function for the predictors ("std" with mean and standard deviation, "robust" with median and interquartile range, and "minmax" with values scaled from 0 to 1). For SLP, the optimal number of neurons (5, 10, 20, 30 and 40) was also included in the grid search. For MLP, the number of hidden layers and neurons were both included simultaneously in the grid search with six possible combinations : 1) two (hidden) layers of 15 and 10 neurons, 2) two layers of 30 and 20 neurons, 3) three layers of 20, 15 and 10 neurons, 4) three layers of 40, 30 and 20 neurons, 5) four layers of 20, 20, 15 and 10 neurons and, 6) four layers of 40, 30, 20 and 20 neurons. Optimal hyperparameters were found by a 5-fold cross-validation on the training set to minimize out-of-sample RMSE. SLP and MLP were fitted using *scikit-learn* in Python (Pedregosa et al., 2011).

For LSTM, the hyperparameters tuning included the learning rate (0.0001, 0.001 and 0.01), the activation function (ReLU and tanh), the addition of a dropout layer of 20% to avoid overfitting, the number of epochs (up to 10 000, with early stopping at 500, 1000, 2500 and 5000 epochs) and the number of cells/layers with five combinations : 1) one layer of 5 cells, 2) 10 cells, 3) 15 cells, 4) two layers of 10 and 5 cells and 5) 15 and 10 cells. Optimal hyperparameters were found to minimize the RMSE on a validation dataset that consisted of 30% of the last years of the training dataset. LSTM was fitted using *keras* in Python (Gulli & Pal, 2017).

FI for neural networks (SLP, MLP and LSTM) were all computed using a permutation-based metric. The mean decrease in RMSE when a predictor was randomly shuffled was computed for each predictor and repeated 200 times. The more the RMSE decreased, the more the predictor was important to the model. FI was computed on both the training and test datasets and scaled to a maximum value of 1.

2.4. Statistical models

Machine and deep learning methods were compared to three statistical models also used in the field: **Generalized Linear Model** (GLM), **Generalized Additive Model** (GAM) and **Distributed Lag Non-Linear Model** (DLNM). GLM considers linear relations between a transformation of the response variable and predictors, while allowing for non-gaussian residuals (Nelder & Wedderburn, 1972). GAM extends GLM with smooth non-linear transformations of the predictors, while keeping the property of additivity of each individual effect (Hastie & Tibshirani, 2017). DLNM describes the exposure-response relationship with a non-linear cross-basis function of lags and exposure variable

(Armstrong, 2006; Gasparrini et al., 2010). In DLNM, a cross-basis function of mean temperature up to 7 days was considered with the same settings as in Gasparrini et al. (2010). Other predictors were also included as non-linear effects as in GAM. All predictors in the statistical models were considered independently (i.e., without interactions), as our goal was to compare models that easily account for potential interactions (machine and deep learning) with models that generally do not (statistical). The statistical models were fitted in R using packages *mgcv* (Wood, 2015) for GLM and GAM and *dlnm* (Gasparrini, 2011) for DLNM. To compare FI metrics of machine/deep learning models with results of statistical models, FI for statistical models was also quantified by the increase of the residual sum of square when one predictor was removed from the model.

2.5. Models' evaluation

Models' performance was assessed using the test dataset (i.e., 30% of observations that were removed before hyperparameters selection and models fitting). Three performance metrics were computed: coefficient of determination (R^2), mean absolute error (MAE) and root mean square error (RMSE). A high value of R^2 is preferred while low values of MAE and RMSE are desired. Note that because R^2 is computed for out-of-sample (test) predictions, its value can be lower than 0, meaning that the model performed worse than a non-informative model (i.e., a model containing only an intercept). In addition to classical performance metrics, a partial validation was also performed. Predictions were compared visually to the observed mortality for the last 10 years, during which two major heatwaves occurred in the CMA of Montreal in 2010 (Bustinza et al., 2013) and 2018 (Lebel et al., 2019). This validation was added to see how the models performed during

extreme heat events of interest lasting several days. Hence, daily observations and predictions were converted to weekly values, which also reduced the noise in the data for this visual assessment. Performance of the models was assessed for models fitted on the main dataset (22 years of data with all predictors), as well for models fitted on the supplementary dataset (39 years of data, without $PM_{2.5}$).

3. Results

The results are only presented for the main dataset (1998–2019, with all predictors) in sections 3.1, 3.2 and 3.3. Then, models' comparison and performance are shown for the models fitted on both datasets (main and supplementary) in sections 3.4 and 3.5.

3.1. Tree-based methods

For DT, the optimal tree depth was found to be 5 in cross-validation (Figure S1). The resulting pruned DT used only three predictors: mean temperature at lag 0, mean NO_2 concentration at lags 1–3 days and mean temperature at lags 1–3 days (Figure 2). Higher temperature and higher NO_2 concentration were both associated with increased summer mortality. For example, temperature above 22ºC (at lag 0) and 26.9ºC (mean of values at lags 1 to 3 days) were found as breakpoints in the heat-mortality relationship (Figure 2).

For RF, the grid search of hyperparameters led to an optimal RF using 5000 trees and the square root of the number of predictors for each tree split (Figure S2). Interestingly, all combinations of hyperparameters for RF led to very close out-of-sample RMSE values of 8.74–8.78 (Figure S2). The two Feature Importance (FI) metrics for RF, based on node purity and permutation, showed that mean temperature, mean temperature at lags 1–3

days and NO_2 concentration at lags 1–3 days were the top 3 predictors (Figure 3a). The fourth most important predictor was either SO_2 or $PM_{2.5}$ concentration (at lags 1–3 days), depending on the FI metric examined.

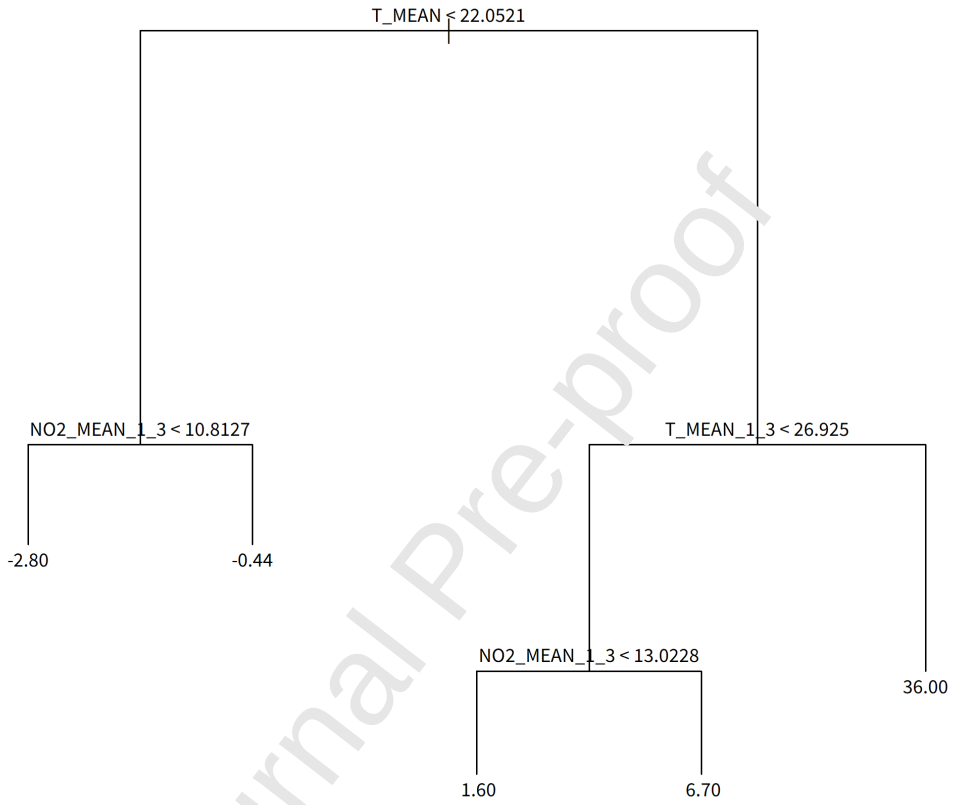


Figure 2: Resulting pruned Decision Tree (DT). The figure shows all predictors' splits (e.g., mean temperature <22.05°C, top of the figure) as well as the predicted daily mortality deviation at each terminal node (e.g., daily mortality deviation = -2.80 when mean temperature is <22.05 and mean NO_2 at lags 1-3 days is <10.81, bottom left of the figure).

Finally, hyperparameters for GBM were also found using the grid search method. The best GBM had a tree depth of 5, 2500 trees, 1/3 of the predictors used at each split and a learning rate of 0.001 (Figure S3). FI showed that the three most important variables were mean temperature, mean temperature at lags 1–3 days and NO_2 concentration at lags 1–3

days (Figure 3b). These most important predictors were the same as in RF, but much further away from the others compared to RF. In GBM, the fourth most important predictor was relative humidity at lags 1–3 days.

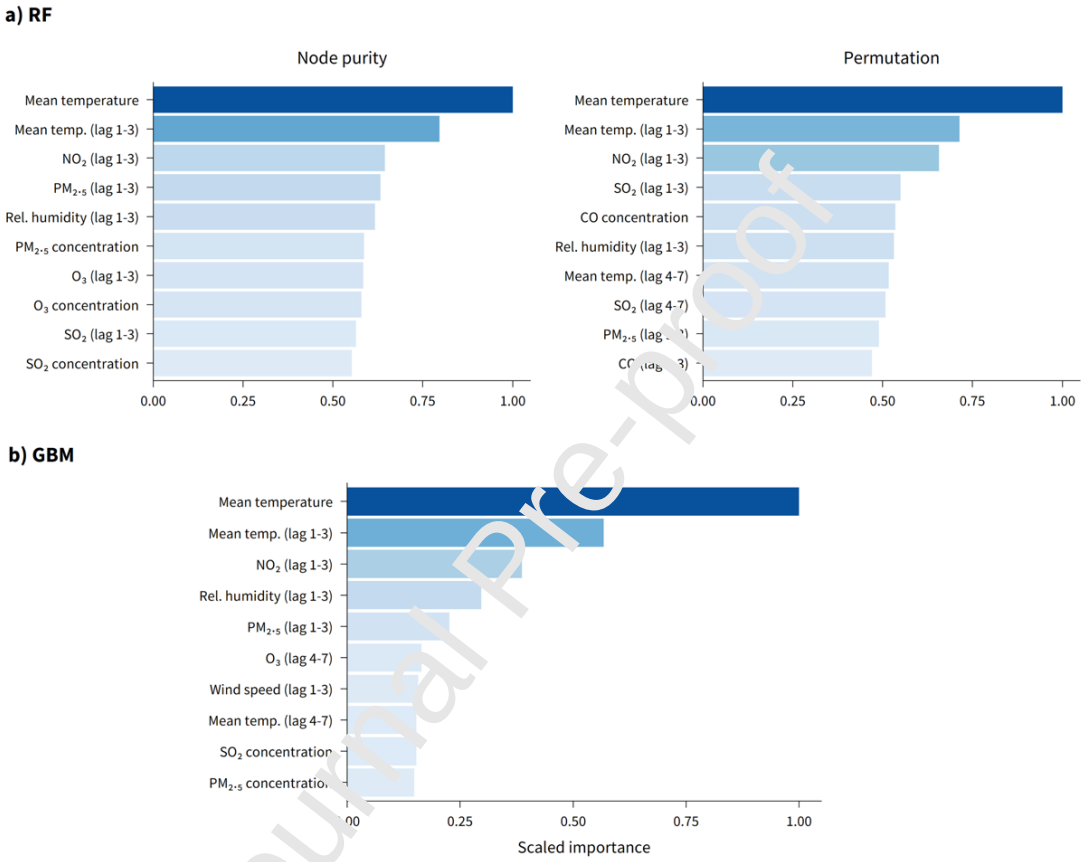


Figure 3: Feature importance (FI) metrics for a) Random Forest (RF) and b) Gradient Booting Machine (GBM). Only the 10 most important predictors are shown for each model.

3.2. Neural networks

For SLP, all models performed better when the "minmax" scaling function was applied to the predictors, compared to the two other methods tested i.e., standard and robust (Figure S4). The SLP with the smallest number of neurons (5) was selected as the best model, using ReLU as an activation function and trained with a learning rate of 0.001

(Figure S4). The FI metrics showed that only one variable seemed to contribute to the predictive power: mean temperature at lag 0 day (Figure 4a). Other important variables, such as relative humidity (at lags 1–3 days) and NO_2 (at lags 1–3 days), had much lower importance in the model. This can be explained by the fact that the optimal SLP found by cross-validation had only 5 neurons, limiting the amount of information that could be learned in that model.

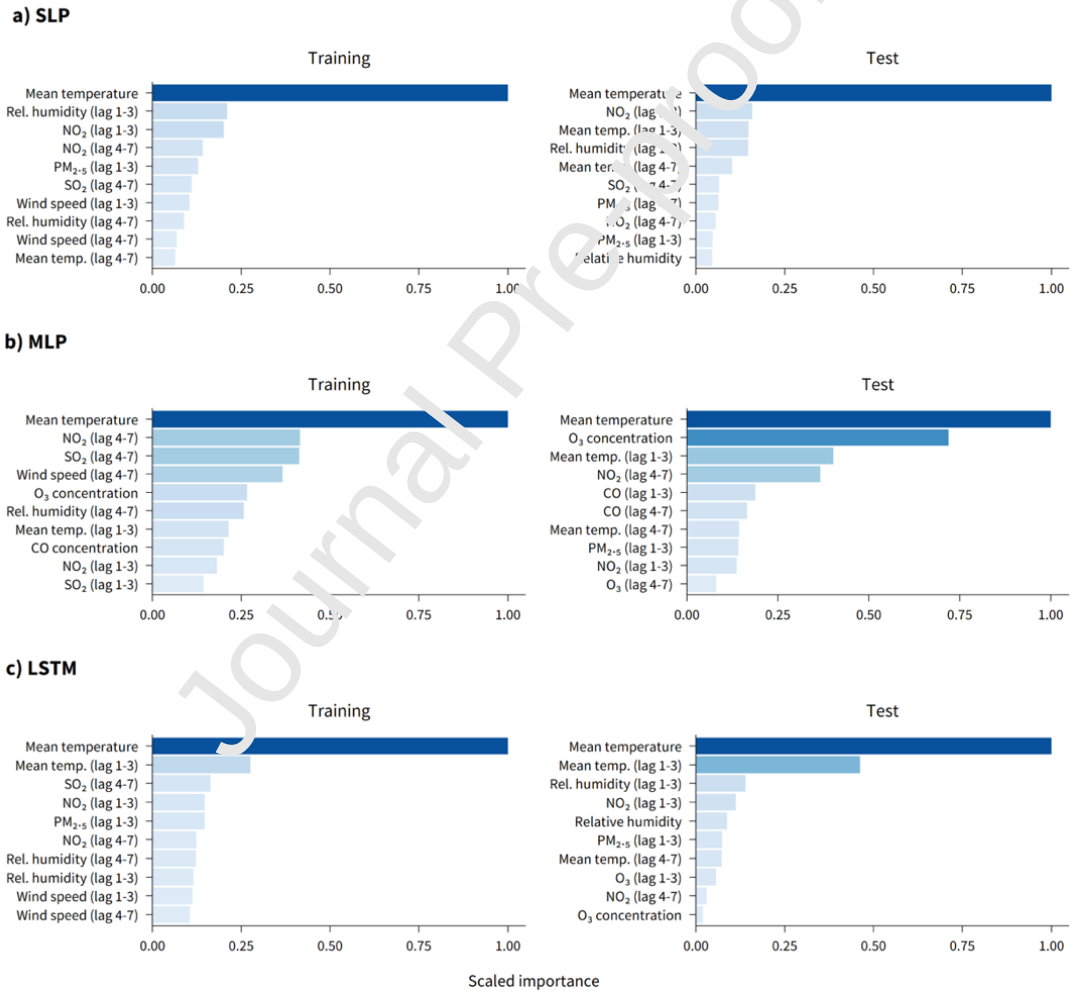


Figure 4: Feature importance (FI) metrics for a) Single-Layer Perceptron (SLP), b) Multi-Layer Perceptron (MLP) and c) Long Short-Term Memory (LSTM). FI metrics are computed for both the training (left) and test (right) datasets. Only the 10 most important predictors are shown for each model.

Like SLP, MLP also performed better when the predictors were scaled using the "minmax" scaler (Figure S5). When two or more hidden layers were considered in MLP, a relatively bigger neural network than SLP was found to be the best by the grid search method. This optimal MLP consisted of three hidden layers with 40 neurons in the first one, 30 in the second one and 20 in the third one. The activation function was logistic and the learning rate was 0.01 (Figure S5). FI metrics showed that, again, mean temperature at lag 0 day was the most important variable (Figure 4b). Results for the other most important predictors differ depending on whether FI was calculated on the training or test dataset. Among the most important variables in MLP, NO_2 and SO_2 (at lags 4–7 days), O_3 concentration (at lag 0 day) and mean temperature (at lags 1–3 days) were found to be relevant contributors.

For the LSTM recurrent neural network, the best model was found among 450 potential candidates (i.e., all combinations of tested hyperparameters). This LSTM consisted of 15 cells in the first hidden layer and 10 cells in the second one and was trained with 10 000 epochs, a ReLU activation function, a learning rate of 0.0001 and no dropout layer (Figure S6). FI metrics for LSTM showed that mean temperature (at lag 0 day) was the most important variable, followed by mean temperature (at lags 1–3 days) for both metrics (Figure 4c). The third most important predictor was either relative humidity (at lags 1–3 days) or SO_2 concentration (at lags 4–7 days) depending on the training or test dataset.

3.3. Statistical models

Top five most important regression coefficients from GLM, GAM and DLNM were extracted and ranked in order of feature importance from left to right (Figure S7). For GLM, the results showed that mean temperature (at lag 0 day) was the most significant variable with a positive association with mortality deviation (Figure S7a). Relative humidity (at lags 1–3 days) was the second most important predictor and had a negative relation to mortality deviation. The third and fourth variables were wind speed (at lags 1–3 days) and mean temperature (at lags 1–3 days). For GAM, the two most important predictors were mean temperature at lags 0 and 1–3 days (Figure S7b). These two temperature variables exhibited the classical U/J shape of temperature-health relationships, which could not be obtained by the FI metrics of tree-based or neural networks models. The third and fourth most important predictors were SO_2 concentration (at lags 4–7 days), that had a negative association with mortality, and NO_2 (at lags 1–3 days), that had a positive one. For DLNM, the bidimensional cross-basis function of air temperature and lags showed a strong positive relationship at lag 0, as well as at lags 1 and 2 days, specifically for high temperature values (Figure S7c). This function that describes the whole temperature-lags-mortality relationship was the most important predictor in the model. The second and third most important predictors were NO_2 (at lags 1–3 days) and SO_2 (at lags 4–7 days) that had similar effects than the ones noted above for GAM.

3.4. Models' comparison

When models using the main dataset were compared in terms of FI, all models agreed that mean temperature (at lag 0 day) was the most important variable to explain mortality deviation (Table 2). In the second rank, mean temperature (at lags 1–3 days) was the most important predictor for all tree-based methods (DT, RF and GBM), LSTM and GAM. Most models also identified NO_2 has a key variable for modelling mortality deviation during summer months, especially at lags 1 to 3 days (for most models) and lags 4 to 7 days (for MLP). Relative humidity also appeared in the second, third or fourth rank for four models (RF, GBM, SLP and GLM) and in the fifth rank for two models (LSTM and GAM). Wind speed was rarely in the top 5 of the most important features except in GLM and DLNM. Air pollutants other than NO_2, such as SO_2, $PM_{2.5}$ and O_3, sometimes appeared in the third to fifth rank of most important predictors.

Table 2: Five most important predictors in all considered models. Lags (in days) are indicated in parentheses (no indication means lag 0). Temperature variables are in light orange, NO_2 in blue and relative humidity in yellow.

		Variable #1	Variable #2	Variable #3	Variable #4	Variable #5
Tree-based methods	DT	Mean temperature	Mean temp. (1-3)	NO_2 (1-3)	-	-
	RF	Mean temperature	Mean temp. (1-3)	NO_2 (1-3)	Rel. hum. (1-3)	SO_2 (1-3)
	GBM	Mean temperature	Mean temp. (1-3)	NO_2 (1-3)	Rel. hum. (1-3)	$PM_{2.5}$ (1-3)
Neural net-works	SLP	Mean temperature	NO_2 (1-3)	Rel. hum. (1-3)	NO_2 (4-7)	SO_2 (4-7)
	MLP	Mean temperature	NO_2 (4-7)	O_3 concentration	Mean temp. (1-3)	NO_2 (1-3)
	LSTM	Mean temperature	Mean temp. (1-3)	NO_2 (1-3)	$PM_{2.5}$ (1-3)	Rel. hum. (1-3)
Statis-tical models	GLM	Mean temperature	Rel. hum. (1-3)	Wind speed (1-3)	Mean temp. (1-3)	NO_2 (1-3)
	GAM	Mean temperature	Mean temp. (1-3)	SO_2 (4-7)	NO_2 (1-3)	Rel. hum. (1-3)
	DLNM	Mean temp. CB	NO_2 (1-3)	SO_2 (4-7)	O_3 (4-7)	Wind speed (1-3)

When FI were extracted for models fitted on the supplementary dataset (that is, the dataset of 39 years of data, but without $PM_{2.5}$), the main conclusions as noted above still held (Table S2). Mean temperature was again the most importance variable in all models. Mean temperature (at lags 1–3 days) was the second most important predictor for 7 out of 9 models. NO_2 (at lags 1–3 days) was consistently the second or third most important predictor, followed closely by relative humidity (at lags 1–3 days). For the other most important variables, the results differed slightly than what was obtained above on the main dataset. For example, SO_2 never appeared in the five most important predictor while wind speed and O_3 concentration came up more often. Finally, mean temperature at lags 4 to 7 days appeared in fifth position for 3 models (GBM, MLP and LSTM), which was not seen when models were trained on the main dataset.

3.5. Models' performance

The ability of the models to predict daily mortality deviation (i.e., over- and under-mortality when long-term and seasonal trends were first removed) on the out-of-sample test set was compared (Figure 5, Table S3a). R^2 did not reach 5% for all models considered and was even negative for some models. These results demonstrated that the studied environmental factors such as temperature-related variables and air pollution only explained a little proportion of the variation in daily mortality. Based on RMSE, MAE and R^2, GBM was the most performing model, closely followed by RF in second rank, especially for the MAE criteria. Then, DT, GAM and DLNM performed equally to model mortality deviation. Only a little performance difference was noted between GAM and DLNM, even though DLNM was given the complete temperature values over the last 7 days, while GAM only used aggregated values of the predictors at lags 0, 1–3, and 4–7

days. For neural networks (SLP, MLP and LSTM), all models performed poorly to model mortality deviation during summer months compared to a non-informative model (i.e., R^2 values close to 0). Finally, GLM had the worst overall results on this dataset.

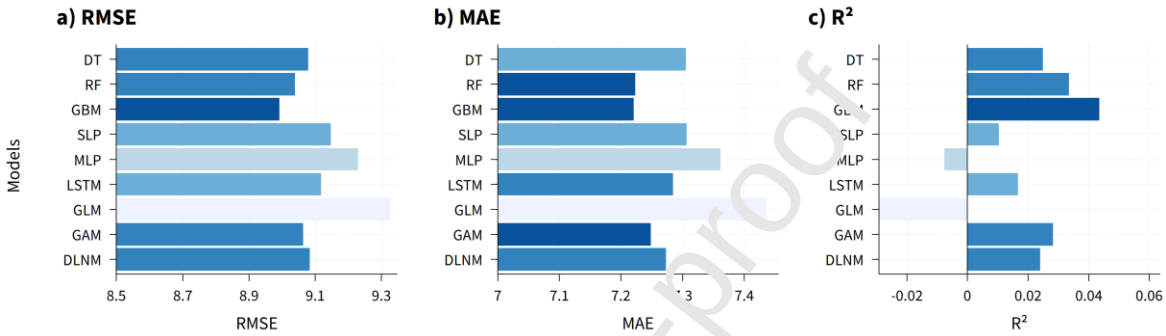


Figure 5 : Performance of all models on the test set (2014–2019) given by a) Root Mean Square Error (RMSE), b) Mean Absolute Error (MAE) and c) Coefficient of determination (R^2). Low values of RMSE and MAE are desired, while high values of R^2 are preferred. The best performance metrics are in dark blue, while the worst are in light blue.

When performance of models fitted using the supplementary dataset was compared, main conclusions held, but some differences were also noted (Figure S8, Table S3b). First, both GBM and RF were again the two best approaches to model mortality deviation. Overall, the performance metrics were better than when fitted on a smaller number of years (best out-of-sample R^2 values of 6.5%), but such comparison should be made with caution given that the test set was not the same (2014–2019 for the main dataset and 2009–2019 for the supplementary dataset). Both MLP and SLP obtained better results compared to when fitted with fewer data, especially MLP, which could mean that the greater size of the dataset helped the MLP to be better calibrated. Overall, LSTM and DT

were the worst models. As highlighted above, GAM outperformed DLNM again. For this dataset, GLM had better performance metrics than GAM and DLNM.

As a final partial validation of the models, weekly mortality predictions for the 2010-2019 period were compared to the observed values, especially for the 2010 and 2018 heatwaves (Figure 6). For the 2010 heatwave, 6 models predicted with great accuracy the over-mortality during this extreme heat event, namely DT, RF, MLP, LSTM, GAM and DLNM. However, since this heatwave was part of the training dataset, no conclusion can be drawn about the performance of the models. For the 2018 heatwave, which was in the out-of-sample test set, only 3 out of the 9 models correctly predicted the peak of mortality, namely DT, GAM and DLNM. These models are relatively simpler models than RF, GBM or neural networks. During more modest heatwaves (e.g., 2011 and 2013), the peak was over-estimated in 2011, but well predicted in 2013, for most models. For non-heatwave years (e.g., 2019), the weekly mortality deviation was generally not well predicted using only temperature-related and air pollution variables.

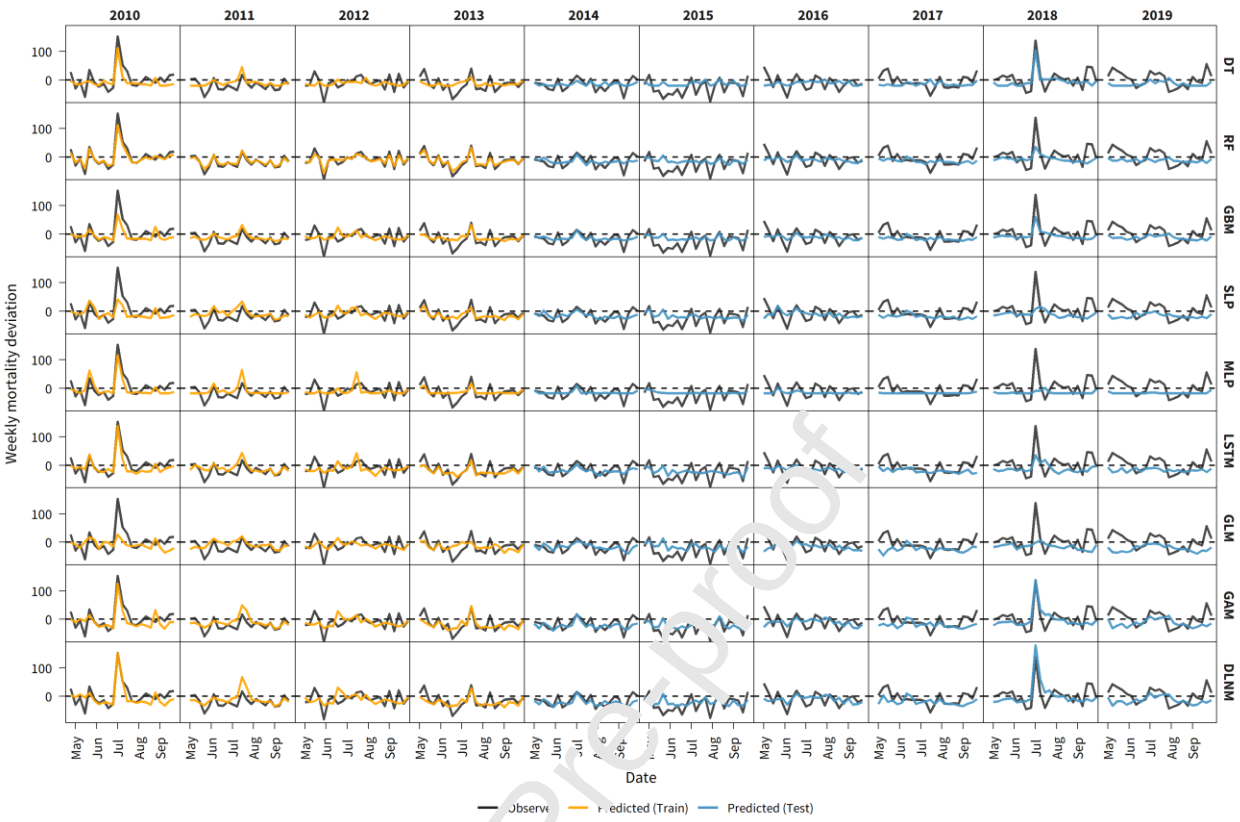


Figure 6: Weekly mortality deviation predictions from all models and observations for the 2010–2019 period.

The same validation was performed with models calibrated using the supplementary dataset (calibration performed from 1991 to 2008), for which both 2010 and 2018 heatwaves were in the out-of-sample test set (Figure S9). Models that could reproduce the over-mortality in 2010 also reproduced appropriately the over-mortality in 2018. These models were the same as noted above for the 2018 heatwave (DT, GAM and DLNM), as well as two other models, RF and GBM.

4. Discussion

This study considered various machine, deep and statistical learning approaches to model the heat-mortality relationship in the metropolitan area of Montreal, Canada. Two datasets of ~20 and ~40 years of data were considered for the calibration of the models. These are much longer than datasets used in the recent literature of <10 years to calibrate such models (e.g., Khatri & Tamil, 2017; Ogata et al., 2021; Park et al., 2020; Qiu et al., 2020; Wang et al., 2019). The hyperparameters optimization results were thoroughly detailed. This contrasts with the existing literature in which hyperparameters are not always indicated (Morgenstern et al., 2020) or optimized (e.g., Zhang et al., 2014), which could lead to a "black box" feeling of the methodology. This new knowledge about the optimal and transparent tuning of machine and deep learning models can facilitate the adoption of these approaches in heat-health studies, as well as increase confidence in these methods by non-expert users.

To emphasize the importance of transparency in machine learning, we made sure that we could explain all models. Indeed, machine/deep learning models are often called black boxes because end users are not provided with information about the contribution of each predictor (Wiemken & Kelley, 2019). Feature Importance (FI) metrics were computed and revealed interesting information in addition to reaffirming established knowledge. The two most important predictors to model daily summer mortality were mean air temperature values at lags 0 and 1 to 3 days, while temperature at lags 4 to 7 days was less important in all models. These results are consistent with the literature and confirmed the relevance of short-term lagged mean temperature to model mortality (Son et al., 2019). Indeed, mean temperature was found to be a good compromise between various

temperature indices for the heat-mortality relationship in comparative studies conducted in the United States (Barnett et al., 2010) and Australia (Vaneckova et al., 2011). NO_2 at lags 1–3 days was overall the third most important variable. This finding is interesting given that NO_2 is less often included in heat-mortality studies that focused more on other air pollutants such as O_3 or PM (Basu, 2009; Son et al., 2019), although it is a known driver of mortality (Y. Wang et al., 2019). The fourth most important predictor was relative humidity at lags 1–3 days. This was expected given the additional effect of humidity on mortality during hot days (Gosling et al., 2009). Although relative humidity is the most commonly used humidity variable in weather-health studies and has been useful in our study for modelling summer mortality using machine and deep learning, it may not be the optimal variable to be considered (Davis et al., 2016). All the above findings were also confirmed when the models were calibrated on the supplementary dataset.

When the ability of the models to predict out-of-sample daily mortality deviation was compared using performance metrics (i.e., RMSE, MAE and R^2), ensemble tree-based methods (GBM and RF) exceeded the performance of traditional statistical models (GLM, GAM and DLNM) and neural networks (SLP, MLP and LSTM). Non-linear statistical models (GAM and DLNM) performed better than neural networks (SLP, MLP and LSTM) when fitted on the 22-year dataset, but MLP had better performance (close to the one of RF) when fitted on the 39-year dataset. This means that a larger dataset size can lead to better performance for more complex models such as neural networks. That said, LSTM performed poorly with both datasets. This could be due to the higher number of parameters to be estimated compared to the other models. Results of GLM and DT

were highly dependent on the dataset (i.e., main or supplementary), so no clear conclusion could be drawn for either model. The best models (GBM and RF) had out-of-sample R^2 values of 3–4% for the main dataset and 6–7% for the supplementary dataset. Such results are not surprising given that R^2 (or explained deviance) below 10% are reported for in-sample predictions in weather-health studies (Table 07 in Chiu, 2017).

As an additional validation, performance of the models during two major heatwaves (i.e., 2010 and 2018) was visually evaluated. Five models (DT, GAM, DLNM, RF and GBM) correctly reproduced the excess mortality during out-of-sample heatwaves when fitted on the 39-year period, but only three (DT, GAM, DLNM) when fitted on the shorter period. Interestingly, models that correctly modelled peak mortality during heatwave were not necessarily the same as the ones having better performance metrics. This could be due in part to the J- or U- shaped relationships between temperature and mortality (Gosling et al., 2009), which were correctly modelled by non-linear statistical models (GAM and DLNM). These results raise the question of the type of validation that should be performed depending on the end user's objective: to predict daily mortality deviations or to reproduce mortality peaks during extreme heat events. In the latter case, partial validation might be more informative for the final model choice than classical performance metrics (e.g., RMSE, MAE, R^2).

To our knowledge, this is the first study to compare such a large variety of approaches to model heat-health relationships, from more easily explainable models to more complex ones. While simpler models can straightforwardly give an indication of the sign of the relationship (GLM, GAM, DLNM) or on potential interactions (DT), more advanced models that allow for complex interactions between predictors may perform better (e.g.,

GBM and RF). We considered 9 modelling approaches, whereas the few other comparative studies in the literature were limited to 3–5 models (e.g., Marien et al., 2022; Nishimura et al., 2021; Ogata et al., 2021; Park et al., 2020; Qiu et al., 2020). Also, it is one of the first applications of the LSTM in the field (e.g., Lin et al., 2021; Nishimura et al., 2021) and a first comparison of the DLNM with machine and deep learning models.

The few other comparative studies found in the literature differ significantly from ours in terms of the studied health outcome, the considered predictors and the models used. For example, Marien et al. (2022) used DT, RF, GBM, MLP and Ridge Regression to model the annual myocardial infarctions and found that MLP and Ridge Regression slightly outperform tree-based methods. Nishimura et al. (2021) modelled the daily number of heat-related-illness using non-linear regression equations, LSTM and RF. Both LSTM and non-linear regression equations outperformed RF. Ogata et al. (2021) compared GLM, GAM, RF and GBM to predict heatstroke and found that GAM was the best model for out-of-sample predictions. M. Park et al. (2020) found that RF was the most performing approach compared to GLM, DT and Support Vector Machine (SVM) to model weekly morbidity due to heatwave. Qiu et al. (2020) modelled days with high hospital admissions for cardiovascular disease and found that RF and GBM performed better than SVM, GLM and DT. These divergent results highlight the need for more comparative studies that consider various modelling approaches and study different health variables (e.g., all-cause mortality in our case).

The main strengths of the study should be highlighted. First, it considered a wide variety of modelling techniques, from statistical models to tree-based methods, to feedforward and recurrent neural networks, thus allowing for complex interactions between

temperature-related and air pollution predictors to be considered. Second, the calibration of the models (i.e., hyperparameters optimization) was transparent and used two long datasets of ~20 and ~40 years of data. Third, FI metrics were extracted and compared for all considered models. They allowed explaining machine and deep learning models and finding the most relevant predictors for heat-health relationships modelling. Finally, two evaluations of models' predictions were performed using three performance metrics and a partial validation based on recent heatwaves.

Some limitations of the study must also be noted. First, only one case study was presented, i.e., the all-cause mortality in the Montreal metropolitan area. Hence, results cannot be directly applied to other regions or health impacts (e.g., cause-specific mortality or morbidity). Second, all models considered the same predictors i.e., mean daily values of temperature-related and air pollution variables at fixed lags 0, 1–3 and 4–7 days. No other lags (except for temperature in DLNM), aggregation (e.g., minimum or maximum), nor temperature metrics (e.g., Humidex, Heat Index) were tested. Third, relative humidity was used in all models considered, but it may not be the best indicator of humidity for heat-health studies. Fourth, FI metrics were computed using either node purity or permutation that are only two of the many methods to explain machine learning models. Finally, even though 9 models were considered, other modelling approaches could have been explored.

5. Conclusion

Mean temperature at lags up to 3 days, as well as NO_2 concentration and relative humidity (both at lags 1–3 days) were the most important predictors for modelling

summer mortality in Montreal, Canada, based on temperature-related and air pollution variables. Ensemble tree-based methods (GBM and RF) outperformed decision tree, statistical models and neural networks based on three performance metrics. However, a partial validation during recent heatwaves showed that these models may underestimate the mortality spike during these events if they are not calibrated with enough data. Therefore, we conclude that both machine/deep learning and statistical models are relevant for modelling heat-health relationships depending on a myriad of factors : size of the dataset, available computing time and resource, information to be derived (e.g., shape of the relationship), end user goal with the fitted model etc. In the context of increased extreme heat events due to climate change, these new results can support the implementation or improvement of heat adaptation measures (e.g., early alert system). Hence, it is suggested that such in-depth comparison of various modelling approaches be extended to other health indicators, predictors and regions.

Acknowledgments

The authors would like to thank Denis Hamel and Louis Rochette (from INSPQ) for their help with the mortality data extraction, and Magalie Canuel, Ray Bustinza, Felix Lamothe, Nathalie Gravel and Yohann Chiu (also from INSPQ) for their comments on early versions of this work. The authors also thank the editor, the associate editor and two anonymous reviewers who helped improve the quality of this paper.

Funding

The main author has received funding from the Natural Sciences and Engineering Research Council of Canada (Vanier Scholarship, #CGV-180821), the Canadian Institute

of Health Research (Health System Impact Fellowship, #IF1-184093), Ouranos (Real-Décoste Excellence Scholarship, #RDX-317725) and the National Institute of Public Health of Quebec (no grant number).

References

Armstrong, B. (2006). Models for the relationship between ambient temperature and daily mortality. *Epidemiology*, 624–631.

Barnett, A., Tong, S., & Clements, A. C. (2010). What measure of temperature is the best predictor of mortality? *Environmental Research*, *110*(6), 604–611.

Basu, R. (2009). High ambient temperature and mortality: A review of epidemiologic studies from 2001 to 2008. *Environmental Health*, *8*(1), 1–13.

Basu, R., Pearson, D., Malig, B., Broadwin, R., & Green, R. (2012). The effect of high ambient temperature on emergency room visits. *Epidemiology*, 813–820.

Basu, R., & Samet, J. M. (2002). Relation between elevated ambient temperature and mortality: A review of the epidemiologic evidence. *Epidemiologic Reviews*, *24*(2), 190–202.

Bayentin, L., El Adlouni, S., Ouarda, T. B., Gosselin, P., Doyon, B., & Chebana, F. (2010). Spatial variability of climate effects on ischemic heart disease hospitalization rates for the period 1989-2006 in Quebec, Canada. *International Journal of Health Geographics*, *9*(1), 1–10.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2).

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Bustinza, R., Lebel, G., Gosselin, P., Bélanger, D., & Chebana, F. (2013). Health impacts of the July 2010 heat wave in Quebec, Canada. *BMC Public Health*, *13*(1), 1–7.

Casati, B., Yagouti, A., & Chaumont, D. (2013). Regional climate projections of extreme heat events in nine pilot Canadian communities for public health planning. *Journal of Applied Meteorology and Climatology*, *52*(12), 2669–2698.

Chiu, Y. M. (2017). *Approches de modélisation des extrêmes dans l'étude des relations entre la santé et la météo*. Université du Québec, Institut national de la recherche scientifique.

Chiu, Y. M., Chebana, F., Abdous, B., Bélanger, D., & Gosselin, P. (2021). Cardiovascular Health Peaks and Meteorological Conditions: A Quantile Regression Approach. *International Journal of Environmental Research and Public Health*, *18*(24). https://doi.org/10.3390/ijerph182413277

Davis, R. E., McGregor, G. R., & Enfield, K. B. (2016). Humidity: A review and primer on atmospheric moisture and human health. *Environmental Research*, *144*, 106–116.

Doyon, B., Bélanger, D., & Gosselin, P. (2008). The potential impact of climate change on annual and seasonal mortality for three cities in Québec, Canada. *International Journal of Health Geographics*, *7*(1), 1–12.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 1189–1232.

Gasparrini, A. (2011). Distributed lag linear and non-linear models in R: the package dlnm. *Journal of Statistical Software*, *43*(8), 1.

Gasparrini, A., Armstrong, B., & Kenward, M. G. (2010). Distributed lag non-linear models. *Statistics in Medicine*, *29*(21), 2224–2234.

Gasparrini, A., Guo, Y., Hashizume, M., Lavigne, E., Zanobetti, A., Schwartz, J., Tobias, A., Tong, S., Rocklöv, J., & Forsberg, B. (2015). Mortality risk attributable to high and low ambient temperature: A multicountry observational study. *The Lancet*, *386*(9991), 369–375.

Gasparrini, A., Guo, Y., Sera, F., Vicedo-Cabrera, A. M., Huber, V., Tong, S., Coelho, M. de S. Z. S., Saldiva, P. H. N., Lavigne, E., & Correa, P. M. (2017). Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health*, *1*(9), e360–e367.

Goldberg, M. S., Gasparrini, A., Armstrong, B., & Valois, M.-F. (2011). The short-term influence of temperature on daily mortality in the temperate climate of Montreal, Canada. *Environmental Research*, *111*(6), 853–860.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Gosling, S. N., Lowe, J. A., McGregor, G. R., Pelling, M., & Malamud, B. D. (2009). Associations between elevated atmospheric temperature and human mortality: A critical

review of the literature. *Climatic Change*, *92*(3), 299–341.

Greenwell, B., Boehmke, B., & Cunningham, J. (2019). Package 'gbm.' *R Package Version*, *2*(5).

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.

Hastie, T., & Tibshirani, R. (2017). *Generalized additive models*. Routledge.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Huang, C., Barnett, A. G., Wang, X., Vaneckova, P., FitzGerald, G., & Tong, S. (2011). Projecting future heat-related mortality under climate change scenarios: A systematic review. *Environmental Health Perspectives*, *119*(12), 1681–1690.

IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*.

Ishigami, A., Hajat, S., Kovats, R. S., Bisanti, L., Rognoni, M., Russo, A., & Paldy, A. (2008). An ecological time-series study of heat-related mortality in three European cities. *Environmental Health*, *7*(1), 1–7.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical*

learning (Vol. 112). Springer.

Jeong, D. I., Sushama, L., Diro, G. T., Khaliq, M. N., Beltrami, H., & Caya, D. (2016). Projected changes to high temperature events for Canada based on a regional climate model ensemble. *Climate Dynamics*, *46*(9), 3163–3180.

Kassomenos, P., Petrakis, M., Sarigiannis, D., Gotti, A., & Karakitsios, S. (2011). Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model. *Air Quality, Atmosphere & Health*, *4*(3), 263–272.

Khatri, K. L., & Tamil, L. S. (2017). Early detection of peak demand days of chronic respiratory diseases emergency department visits using artificial neural networks. *IEEE Journal of Biomedical and Health Informatics*, *22*(1), 285–290.

Kovats, R. S., & Hajat, S. (2008). Heat stress and public health: A critical review. *Annu. Rev. Public Health*, *29*, 41–55.

Lavigne, E., Gasparrini, A., Wang, X., Chen, H., Yagouti, A., Fleury, M. D., & Cakmak, S. (2014). Extreme ambient temperatures and cardiorespiratory emergency room visits: Assessing risk by comorbid health conditions in a time series study. *Environmental Health*, *13*(1), 1–8.

Lebel, G., Dubé, M., & Bustinza, R. (2019). Surveillance des impacts des vagues de chaleur extrême sur la santé au Québec à l'été 2018. *Bulletin d'information En Santé Environnementale*, *1*, 1–10.

Li, M., Gu, S., Bi, P., Yang, J., & Liu, Q. (2015). Heat waves and morbidity: Current knowledge and further direction-a comprehensive literature review. *International Journal of Environmental Research and Public Health*, *12*(5), 5256–5283.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.

Lin, S., Hsu, W.-H., Van Zutphen, A. R., Saha, S., Luber, G., & Hwang, S.-A. (2012). Excessive heat and respiratory hospitalizations in New York State: Estimating current and future public health burden related to climate change. *Environmental Health Perspectives*, *120*(11), 1571–1577.

Lin, Y.-C., Tsai, C.-H., Hsu, H.-T., & Lin, C.-H. (2021). *Using Machine Learning to Analyze and Predict the Relations Between Cardiovascular Disease Incidence, Extreme Temperature and Air Pollution*. 234–237.

Marien, L., Valizadeh, M., Zu Castell, W., Nam, C., Rechid, D., Schneider, A., Meisinger, C., Linseisen, J., Wolf, K., & Bouwer, L. (2022). Machine learning models to predict myocardial infarctions from past climatic and environmental conditions. *Natural Hazards and Earth System Sciences Discussions*, 1–36.

Masselot, P., Chebana, F., Bélanger, D., St-Hilaire, A., Abdous, B., Gosselin, P., & Ouarda, T. B. (2018). Aggregating the response in time series regression models, applied to weather-related cardiovascular mortality. *Science of The Total Environment*, *628*, 217–225.

Masselot, P., Chebana, F., Campagna, C., Lavigne, É., Ouarda, T. B., & Gosselin, P.

(2021). Machine learning approaches to identify thresholds in a heat-health warning system context. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*.

Meehl, G. A., & Tebaldi, C. (2004). More intense, more frequent, and longer lasting heat waves in the 21st century. *Science*, *305*(5686), 994–997.

Mora, C., Dousset, B., Caldwell, I. R., Powell, F. E., Geronimo, R. C., Bielecki, C. R., Counsell, C. W., Dietrich, B. S., Johnston, E. T., & Louis, L. V. (2017). Global risk of deadly heat. *Nature Climate Change*, *7*(7), 501–506.

Morgenstern, J. D., Buajitti, E., O'Neill, M., Piggott, T., Goel, V., Fridman, D., Kornas, K., & Rosella, L. C. (2020). Predicting population health with machine learning: A scoping review. *BMJ Open*, *10*(10), e037863.

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, *135*(3), 370–384.

Nishimura, T., Rashed, E. A., Kodera, S., Shirakami, H., Kawaguchi, R., Watanabe, K., Nemoto, M., & Hirata, A. (2021). Social implementation and intervention with estimated morbidity of heat-related illnesses from weather data: A case study from Nagoya City, Japan. *Sustainable Cities and Society*, *74*, 103203.

Ogata, S., Takegami, M., Ozaki, T., Nakashima, T., Onozuka, D., Murata, S., Nakaoku, Y., Suzuki, K., Hagihara, A., & Noguchi, T. (2021). Heatstroke predictions by machine learning, weather information, and an all-population registry for 12-hour heatstroke alerts. *Nature Communications*, *12*(1), 1–11.

Park, J., & Kim, J. (2018). Defining heatwave thresholds using an inductive machine learning approach. *Plos One*, *13*(11), e0206872.

Park, M., Jung, D., Lee, S., & Park, S. (2020). Heatwave Damage Prediction Using Random Forest Model in Korea. *Applied Sciences*, *10*(22), 8237.

Pascal, M., Goria, S., Wagner, V., Sabastia, M., Guillet, A., Cordeau, E., Mauclair, C., & Host, S. (2021). Greening is a promising but likely insufficient adaptation strategy to limit the health impacts of extreme heat. *Environment International*, *151*, 106441.

Pascal, M., Wagner, V., Corso, M., Laaidi, K., Ung, A., & Beaudeau, P. (2018). Heat and cold related-mortality in 18 French cities. *Environment International*, *121*, 189–198.

Pascal, M., Wagner, V., Le Tertre, A., Laaidi, K., Honoré, C., Bénichou, F., & Beaudeau, P. (2013). Definition of temperature thresholds: The example of the French heat wave warning system. *International Journal of Biometeorology*, *57*(1), 21–29.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Qiu, H., Luo, L., Su, Z., Zhou, L., Wang, L., & Chen, Y. (2020). Machine learning approaches to predict peak demand days of cardiovascular admissions considering environmental exposure. *BMC Medical Informatics and Decision Making*, *20*(1), 1–11.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Schwartz, J., Samet, J. M., & Patz, J. A. (2004). Hospital admissions for heart disease:

The effects of temperature and humidity. *Epidemiology*, *15*(6), 755–761.

Son, J.-Y., Liu, J. C., & Bell, M. L. (2019). Temperature-related mortality: A systematic review and investigation of effect modifiers. *Environmental Research Letters*, *14*(7), 073004.

Statistics Canada. (2022). *Census metropolitan area (CMA) and census agglomeration (CA)*. https://www150.statcan.gc.ca/n1/pub/92-195-x/2021001/geo/cma-rmr/cma-rmr-eng.htm

Tong, S., & Kan, H. (2011). Heatwaves: What is in a definition? *Maturitas*, *69*(1), 5–6.

Vaneckova, P., Neville, G., Tippett, V., Aitken, R., FitzGerald, G., & Tong, S. (2011). Do biometeorological indices improve modelling outcomes of heat-related mortality? *Journal of Applied Meteorology and Climatology*, *50*(6), 1165–1176.

Vicedo-Cabrera, A. M., Scovronick, N., Sera, F., Royé, D., Schneider, R., Tobias, A., Astrom, C., Guo, Y., Honda, Y., & Hondula, D. M. (2021). The burden of heat-related mortality attributable to recent human-induced climate change. *Nature Climate Change*, *11*(6), 492–500.

Vicedo-Cabrera, A. M., Sera, F., Guo, Y., Chung, Y., Arbuthnott, K., Tong, S., Tobias, A., Lavigne, E., Coelho, M. de S. Z. S., & Saldiva, P. H. N. (2018). A multi-country analysis on potential adaptive mechanisms to cold and heat in a changing climate. *Environment International*, *111*, 239–246.

Wang, X., Lavigne, E., Ouellette-kuntz, H., & Chen, B. E. (2014). Acute impacts of

extreme temperature exposure on emergency room admissions related to mental and behavior disorders in Toronto, Canada. *Journal of Affective Disorders*, *155*, 154–161.

Wang, Y., Song, Q., Du, Y., Wang, J., Zhou, J., Du, Z., & Li, T. (2019). A random forest model to predict heatstroke occurrence for heatwave in China. *Science of the Total Environment*, *650*, 3048–3053.

Wiemken, T. L., & Kelley, R. R. (2019). Machine Learning in Epidemiology and Health Outcomes Research. *Annual Review of Public Health*, *41*, 21–36.

Wondmagegn, B. Y., Xiang, J., Williams, S., Pisaniello, D., & Bi, P. (2019). What do we know about the healthcare costs of extreme heat exposure? A comprehensive literature review. *Science of the Total Environment*, *657*, 608–618.

Wood, S. (2015). Package 'mgcv.' *R Package Version*, *1*(29), 729.

Xu, Z., FitzGerald, G., Guo, Y., Jalaludin, B., & Tong, S. (2016). Impact of heatwave on mortality under different heatwave definitions: A systematic review and meta-analysis. *Environment International*, *89*, 193–203.

Ye, X., Wolff, R., Yu, W., Vaneckova, P., Pan, X., & Tong, S. (2012). Ambient temperature and morbidity: A review of epidemiological evidence. *Environmental Health Perspectives*, *120*(1), 19–28.

Zhang, K., Li, Y., & Schwartz, J. D. (2014). What weather variables are important in predicting heat-related mortality? A new application of statistical learning methods. *Environmental Research*, *132*, 350–359.

Zhang, K., Rood, R. B., Michailidis, G., Oswald, E. M., Schwartz, J. D., Zanobetti, A., Ebi, K. L., & O'Neill, M. S. (2012). Comparing exposure metrics for classifying 'dangerous heat'in heat wave and health warning systems. *Environment International*, *46*, 23–29.

CRediT authorship contribution statement

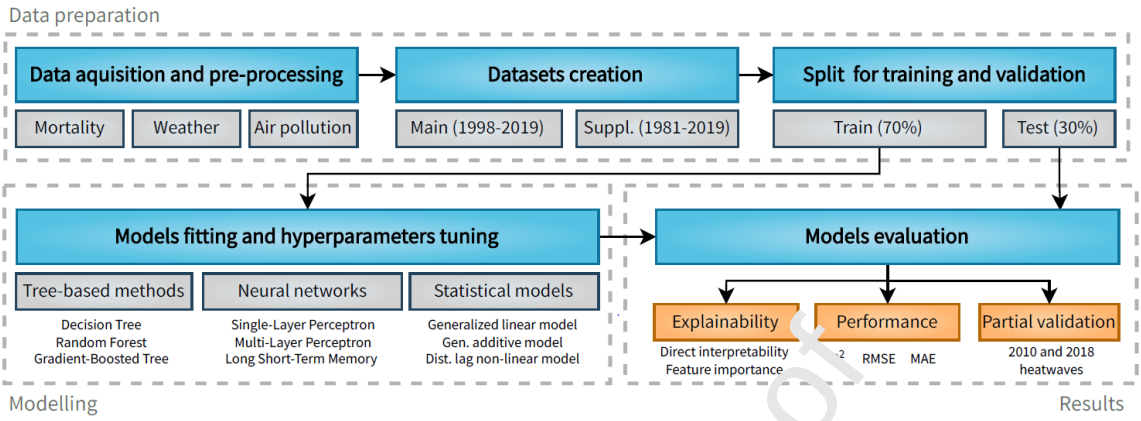
Jeremie Boudreault: Conceptualization, Methodology, Data Curation, Formal analysis, Visualization, Software, Writing - Original Draft, Review and Editing, Funding acquisition. **Celine Campagna**: Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition. **Fateh Chebana**: Conceptualization, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Graphical abstract



Highlights :

- Heat-health relationship modelled with 9 machine/deep/statistical learning models
- Interactions considered between temperature-related and air pollution variables
- Air temperature, NO_2 and relative humidity were the most important predictors
- Ensemble tree-based models outperformed neural networks and statistical models
- Non-linear statistical models may better represent peak mortality during heatwaves

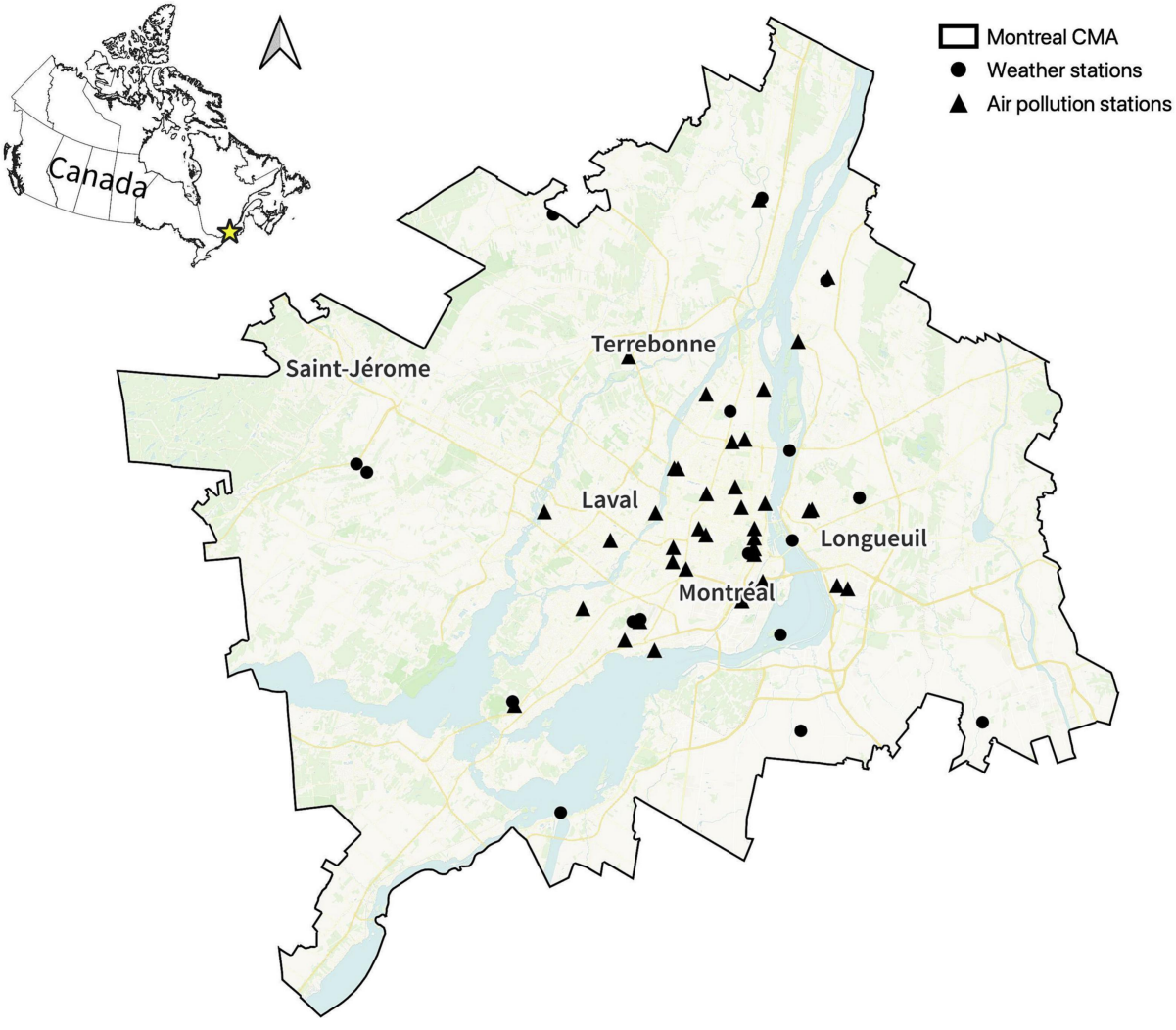

Canada
Montreal CMA
Weather stations
Air pollution stations
Saint-Jérome
Terrebonne
Laval
Longueuil
Montréal

Figure 1

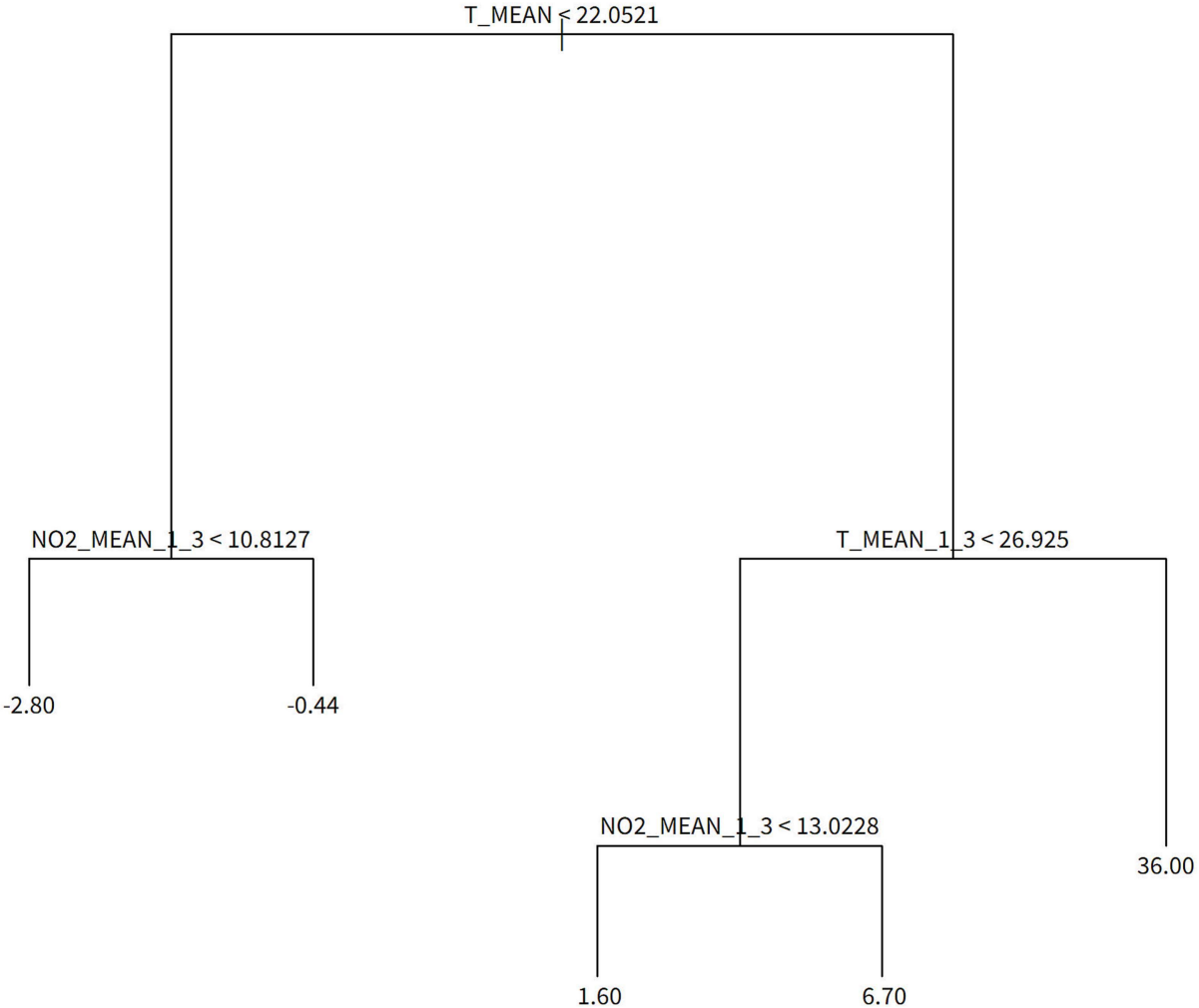

T_MEAN < 22.0521
NO2_MEAN_1_3 < 10.8127
T_MEAN_1_3 < 26.925
-2.80
-0.44
NO2_MEAN_1_3 < 13.0228
36.00
1.60
6.70

Figure 2

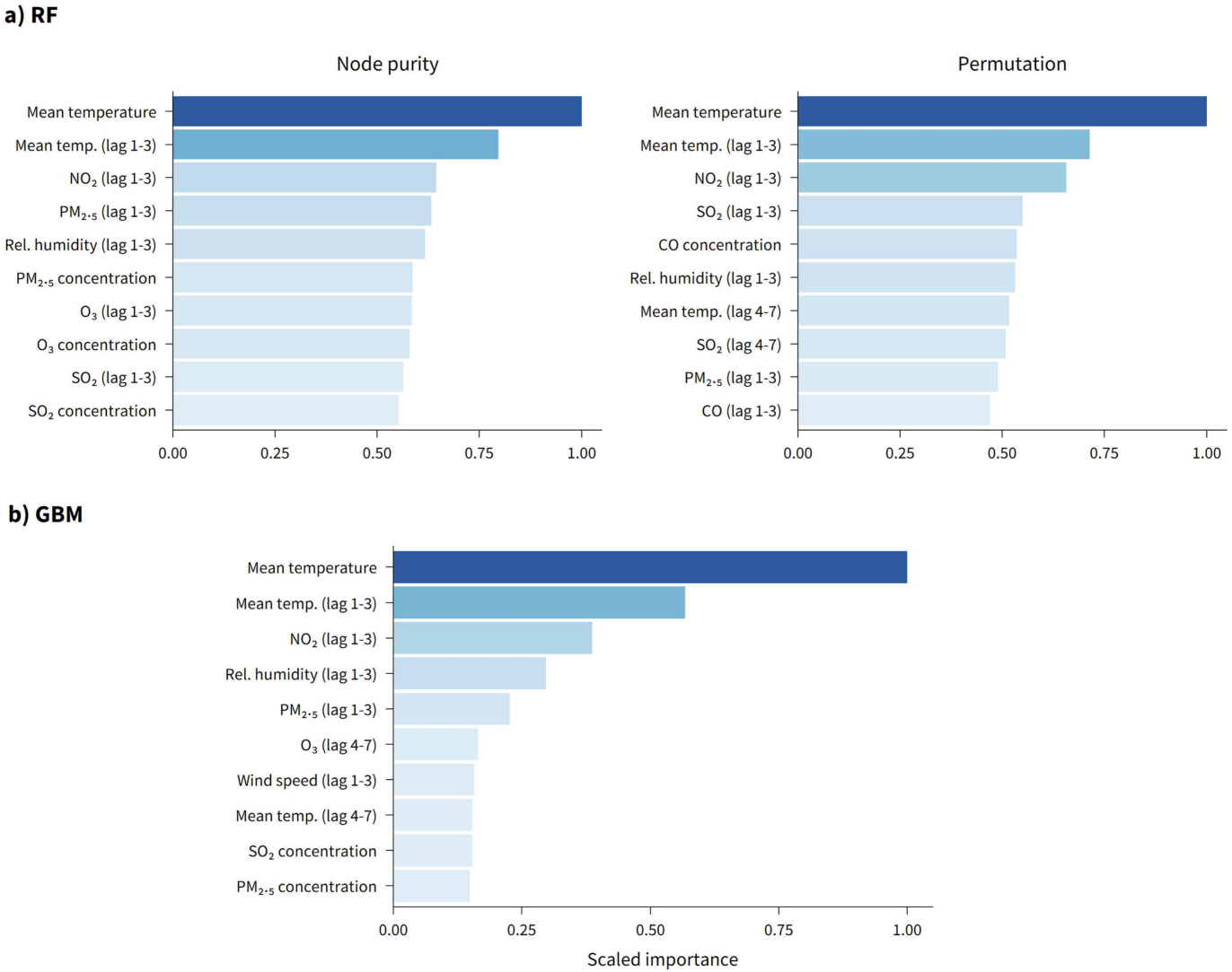

a) RF
Node purity
Mean temperature
Mean temp. (lag 1-3)
NO2 (lag 1-3)
PM2.5 (lag 1-3)
Rel. humidity (lag 1-3)
PM2.5 concentration
O3 (lag 1-3)
O3 concentration
SO2 (lag 1-3)
SO2 concentration
0.00
0.25
0.50
0.75
1.00
Permutation
Mean temperature
Mean temp. (lag 1-3)
NO2 (lag 1-3)
SO2 (lag 1-3)
CO concentration
Rel. humidity (lag 1-3)
Mean temp. (lag 4-7)
SO2 (lag 4-7)
PM2.5 (lag 1-3)
CO (lag 1-3)
0.00
0.25
0.50
0.75
1.00
b) GBM
Mean temperature
Mean temp. (lag 1-3)
NO2 (lag 1-3)
Rel. humidity (lag 1-3)
PM2.5 (lag 1-3)
O3 (lag 4-7)
Wind speed (lag 1-3)
Mean temp. (lag 4-7)
SO2 concentration
PM2.5 concentration
0.00
0.25
0.50
0.75
1.00
Scaled importance

Figure 3

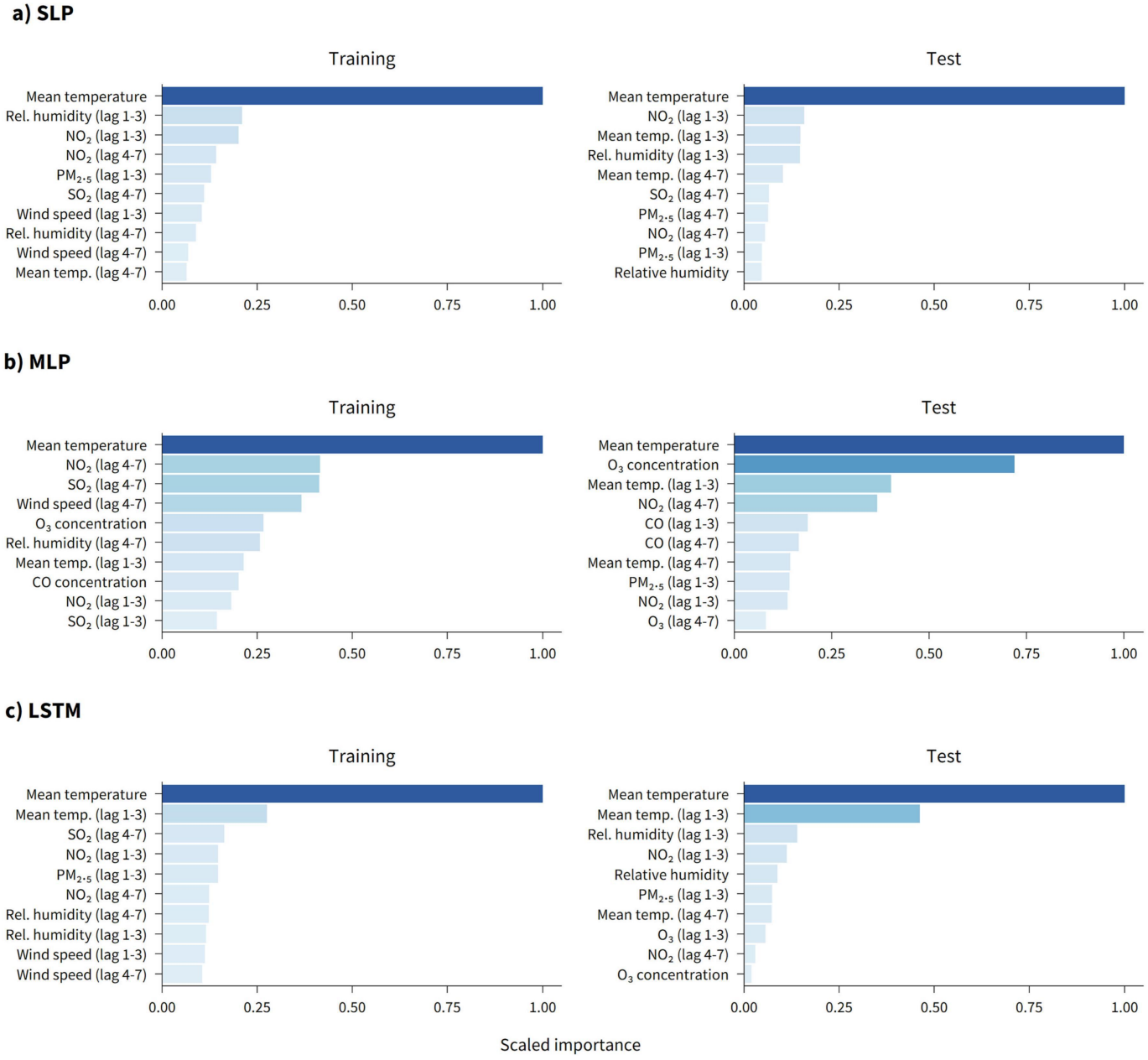

a) SLP
Training
Mean temperature
Rel. humidity (lag 1-3)
NO2 (lag 1-3)
NO2 (lag 4-7)
PM2.5 (lag 1-3)
SO2 (lag 4-7)
Wind speed (lag 1-3)
Rel. humidity (lag 4-7)
Wind speed (lag 4-7)
Mean temp. (lag 4-7)
Test
Mean temperature
NO2 (lag 1-3)
Mean temp. (lag 1-3)
Rel. humidity (lag 1-3)
Mean temp. (lag 4-7)
SO2 (lag 4-7)
PM2.5 (lag 4-7)
NO2 (lag 4-7)
PM2.5 (lag 1-3)
Relative humidity
b) MLP
Training
Mean temperature
NO2 (lag 4-7)
SO2 (lag 4-7)
Wind speed (lag 4-7)
O3 concentration
Rel. humidity (lag 4-7)
Mean temp. (lag 1-3)
CO concentration
NO2 (lag 1-3)
SO2 (lag 1-3)
Test
Mean temperature
O3 concentration
Mean temp. (lag 1-3)
NO2 (lag 4-7)
CO (lag 1-3)
CO (lag 4-7)
Mean temp. (lag 4-7)
PM2.5 (lag 1-3)
NO2 (lag 1-3)
O3 (lag 4-7)
c) LSTM
Training
Mean temperature
Mean temp. (lag 1-3)
SO2 (lag 4-7)
NO2 (lag 1-3)
PM2.5 (lag 1-3)
NO2 (lag 4-7)
Rel. humidity (lag 4-7)
Rel. humidity (lag 1-3)
Wind speed (lag 1-3)
Wind speed (lag 4-7)
Test
Mean temperature
Mean temp. (lag 1-3)
Rel. humidity (lag 1-3)
NO2 (lag 1-3)
Relative humidity
PM2.5 (lag 1-3)
Mean temp. (lag 4-7)
O3 (lag 1-3)
NO2 (lag 4-7)
O3 concentration
0.00
0.25
0.50
0.75
1.00
Scaled importance

Figure 4

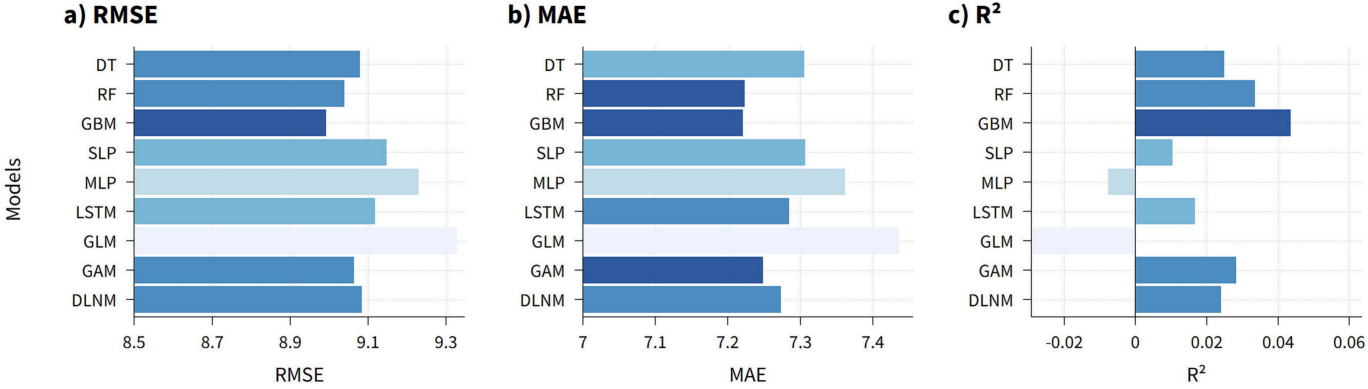

a) RMSE
b) MAE
c) R²
Models
DT
RF
GBM
SLP
MLP
LSTM
GLM
GAM
DLNM
8.5
8.7
8.9
9.1
9.3
RMSE
7
7.1
7.2
7.3
7.4
MAE
-0.02
0
0.02
0.04
0.06
R²

Figure 5

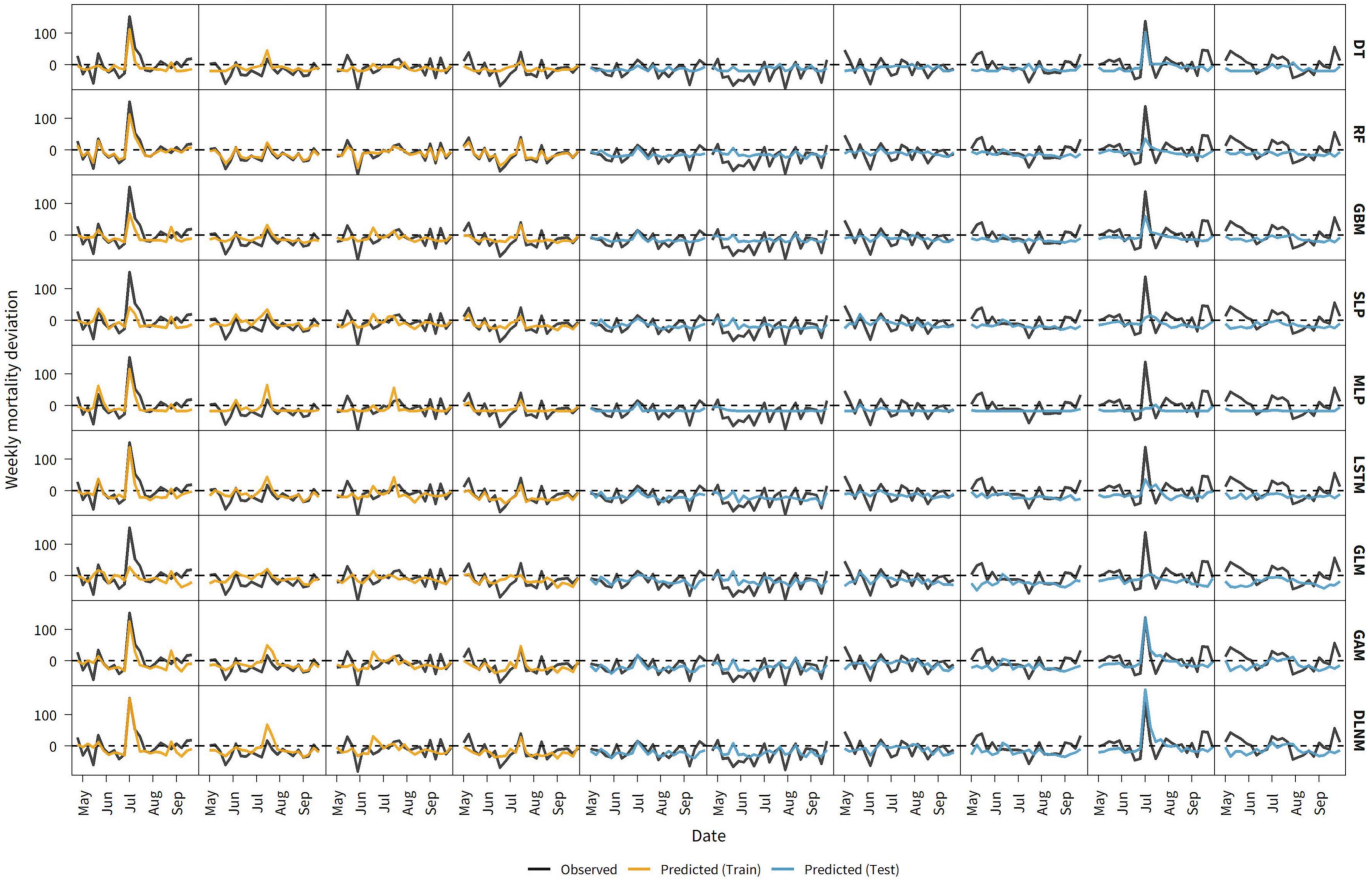

2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
DT
RF
GBM
SLP
MLP
LSTM
GLM
GAM
DLNM
100
0
Weekly mortality deviation
May
Jun
Jul
Aug
Sep
Date
Observed
Predicted (Train)
Predicted (Test)

Figure 6