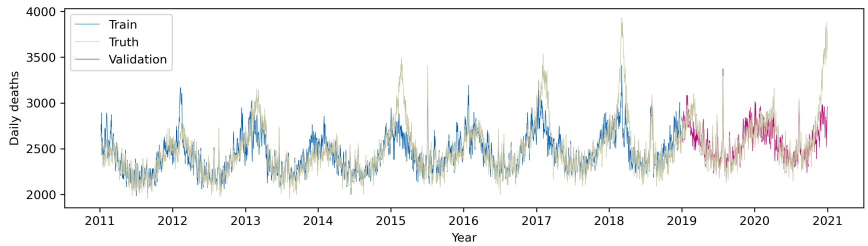
Supplementary 1 Model Performance



Extended Data Figure 1: Registered and estimated daily mortality in Germany with the best model. The temperature data in 2013 was replaced with the interpolated DWD data.

Extended Data Table 1: Performance of best model

RMSE	train	validation
hot days	91.4	83.9
all days	171.1	199.6

The models were trained using the average temperature data from Helmholtz Munich with the average of the 10 lowest weeks in the previous year as mortality baseline. The reported results are averages derived from 20 ensemble members. Root mean square errors were calculated for hot days (days with a Germany-wide average temperature exceeding 20°C) and all days. The correction of day-of-the-week was applied. Data from the year 2013 was excluded due to irregularities.

Supplementary 2 Impact of model parameters

Results

In our paper, we explored the impact of different factors on the accuracy of our model to gain a further insight into the relationship between temperature and mortality.

We implemented two different shallow neural networks: one linear network with rectified linear unit (ReLU) as the activation function, and another utilizing a network with an exponential activation function. The exponential activation function was inspired by the statistical method. To mitigate the effect of the random model initialization, we trained multiple instances as ensembles.

The fully connected network is more susceptible to overfitting (RMSE: 88.9 for training dataset, 136.9 for validation dataset). In contrast, the exponential model demonstrates greater consistency across both the training and the validation datasets (RMSE: 112.9 for training dataset, 107.4 for validation dataset, Extended Data Table 3). Furthermore, the exponential model exhibited remarkable consistency among various trained instances (standard error 1.7 on validation dataset, Extended Data Table 3). Conversely, the fully connected model exhibited notable variations across different trained instances (standard error 8.1 on validation dataset). Therefore, we decided to use the exponential model for further study.

The past studies have demonstrated the lag effect of the temperature on the mortality. We examined the effect of lag days with different 1-d convolutional kernel size (0 to 9 days). The accuracy of the exponential model initially improved as we increased the number of lag days, but reached a plateau after 5 lag days (Extended Data Table 2).

The baseline of the death rate is essential to estimate the number of the death cases and the heat-related mortality. The baselines varies across different age groups and genders, and are also highly dynamic with time. Shkolnikov et al. has discussed the impact of different types of baseline on the excess death estimation[2]. In this paper, the baseline is the mortality with an ideal temperature. We tested 3 different kinds of baseline: a constant baseline, the baseline with the average mortality from the previous year and the baseline with the average mortality from the 10 lowest weeks in the last year. Extended Data Figure 2 illustrated the change of different baselines with time. The results of our model showed that using the average mortality from the 10 lowest weeks in the last year provided an accurate estimate of daily mortality in Germany (RMSE, 96.9 for training dataset, 98.3 for validation dataset). The model with this baseline outperformed the ones using the baseline with the average mortality from the previous year (RMSE, 113.0 for training dataset, 107.1 for validation dataset) or a constant baseline (RMSE, 115.9 for training dataset, 125.4 for validation dataset, Extended Data Table 4).

Our results also showed that the accuracy of the model depends strongly on the weather data. The model trained with district temperature derived from the Copernicus regional reanalysis for Europe (CERRA) dataset, which has a lower resolution, demonstrated relatively poor performance (RMSE, 97.6 for training dataset, 133.5 for validation dataset). Regarding the high resolution weather data, the model trained with mean temperature provided the most accurate estimations, and the model trained with minimum temperature outperformed the one trained with maximum temperature. Using all three temperature variables together as input showed only marginal improvement (RMSE, 94.8 for training dataset, 97.6 for validation dataset, Extended Data Table 5).

The registered death cases showed a significant gap between workdays and weekends. Therefore, we introduced a trainable parameter for each day in a week and used it to adjust the prediction. This correction significantly improved the accuracy (RMSE, 91.4 for training dataset, 83.9 for validation dataset, Extended Data Table 6).

The most accurate model in our study was the exponential model trained with the daily mean temperature from the high resolution weather data with

5 days lag effect, with the average mortality from the 10 lowest weeks in the last year as the baseline for mortality calculation. The best model also included correction factors to tackle the mortality gaps between weekdays and weekends.

Extended Data Table 2: Impact of model structure and lag days

RMSE	ReLU		Exponential	
lag days	train	validation	train	validation
no lag	145.82	174.54	149.44	193.13
1 day	115.17	131.83	123.89	158.33
3 days	106.47	125.71	110.77	133.44
4 days	100.83	137.98	110.39	122.14
5 days	98.98	153.30	111.82	111.72
6 days	95.05	149.04	113.31	107.68
7 days	94.31	140.79	115.89	108.74
8 days	90.70	141.41	114.68	107.55
9 days	86.28	137.93	110.76	109.31
10 days	84.44	140.18	108.00	113.25

The models were trained using the average temperature data from Helmholtz Munich, with the average mortality from the previous year serving as the baseline. Day-of-the-week correction was not included in the training. The reported results are averages derived from 20 ensemble members. Root mean square errors were calculated for hot days, defined as days with a Germany-wide average temperature exceeding 20°C. Data from the year 2013 was excluded due to irregularities.

Extended Data Table 3: Variance within ensemble members

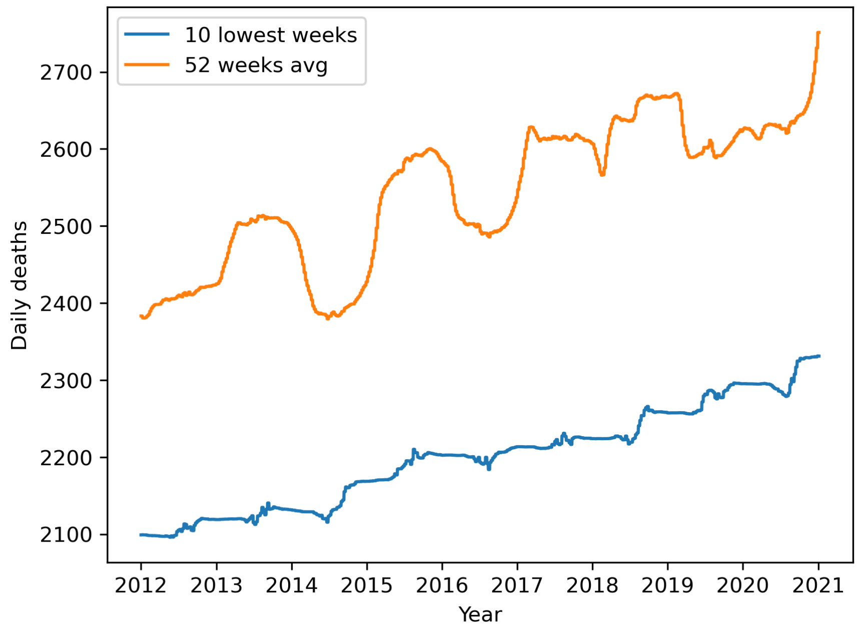
RMSE	ReLU		Exponential	
5 lag days	train	validation	train	validation
error of mean	88.88	136.88	112.94	107.43
mean of error	95.78± 3.25	147.61±8.10	113.38±0.88	108.18±1.65
max of error	103.69	167.06	116.07	113.06

The models were trained using the average temperature data from Helmholtz Munich, with the average mortality from the previous year serving as the baseline. Day-of-the-week correction was not included in the training. The reported results are averages derived from 100 ensemble members. Root mean square errors were calculated for hot days, defined as days with a Germany-wide average temperature exceeding 20°C. We employed two different methods to calculate the RMSE. In the 'error of the mean,' we calculated the average result of all ensemble members and compared it to registered deaths. In the 'mean of error,' we compared the result of each individual ensemble member to registered deaths to obtain the RMSE. We then calculated the corresponding mean and standard error. Additionally, we listed the maximum error among all ensemble members as a reference. Data from the year 2013 was excluded due to irregularities.

Extended Data Table 4: Impact of baseline selection

RMSE	Exponential	
5 lag days	train	validation
constant	115.92	125.36
52 weeks avg	112.99	107.14
10 lowest weeks	96.91	98.28

The models were trained using the average temperature data from Helmholtz Munich with different baselines. Day-of-the-week correction was not included in the training. The reported results are averages derived from 20 ensemble members. Root mean square errors were calculated for hot days, defined as days with a Germany-wide average temperature exceeding 20°C. Data from the year 2013 was excluded due to irregularities.



Extended Data Figure 2: Estimated mortality baseline with two different methods.

Extended Data Table 5: Impact of temperature dataset

RMSE	Exponential	
5 lag days	train	validation
1×1 km, max	117.40	117.68
1×1 km, mean	96.91	98.28
1×1 km, min	110.06	109.07
1×1 km, 3 channels	94.83	97.59
CERRA, 6-hourly	97.57	133.48

The models were trained with the average of the 10 lowest weeks in the previous year as mortality baseline. Day-of-the-week correction was not included in the training. The reported results are averages derived from 20 ensemble members. Root mean square errors were calculated for hot days, defined as days with a Germany-wide average temperature exceeding 20°C. Data from the year 2013 was excluded due to irregularities.

Extended Data Table 6: Impact of day-of-the-week correction

RMSE	Exponential	
5 lag days	train	validation
no correction	96.91	98.28
with correction	91.43	83.85

The models were trained using the average temperature data from Helmholtz Munich with the average of the 10 lowest weeks in the previous year as mortality baseline. The reported results are averages derived from 20 ensemble members. Root mean square errors were calculated for hot days, defined as days with a Germany-wide average temperature exceeding 20°C. Data from the year 2013 was excluded due to irregularities.

Discussions

The daily temperature data used in our model allows for a closer examination of the temperature's influence on mortality. In addition, we tested the effects of maximum, minimum and average temperatures on mortality. While the model using average temperature demonstrated the best performance, the model using minimum temperature outperformed the one using maximum temperature. These results confirmed the findings of previous studies, which suggest that high night temperatures during a heatwave are more fatal to vulnerable populations[1].

Additionally, our model enabled further inspection into the factors impacting the model's accuracy. Shkolnikov et al. suggested different methods for estimating the mortality baseline[2]. The suggested baseline estimated a higher number of excess deaths compared to the traditional method. However, no validation was provided in the paper. With our model, we confirmed that the average mortality from the 10 lowest weeks in the last year is a closer estimation of the mortality rate baseline compared to the traditional methods. The main reason for the differences is the inclusion of seasonal influenza impact in yearly averages The impact leads to strong variation of the baseline and adversely affects the performance of the model. These results further indicated that the conventional excess death estimation underestimated the potentially avoidable losses, especially in winter. It is important to mention that the COVID-19 had an significant impact on mortality baseline estimation. Typically, the 10 weeks with the lowest mortality are in the summer months. However, in 2022, mortality during the summer was unusually high due to the impact of COVID-19, making it unsuitable for the baseline estimation.

The performance improvement with the correction factors for each day of the week highlighted irregularities in the registration data. Death registrations may experience delays during weekends or holidays. Thus, it is challenging to separate the lag effect due to the delay of the registration from the lag effect of the temperature on mortality. As we did not find national-wide hospitalization data, we decided to rely on the data from the official death registration and used different methods to validate our model. Further studies of the relationship between the mortality and hospitalization during heatwaves will help to guide the preparation of healthcare system of future heatwaves.

Supplementary 3 Impact of temperature aggregation

Extended Data Table 7: Estimated heat-related mortality with different temperature aggregation level. The negative numbers are the differences to the estimation with district level temperature.

aggregation level	District	State	Germany
2014	2000	1700 *(-300)*	1100 *(-900)*
2015	7800	6500 *(-1300)*	5200 *(-2600)*
2016	2400	2100 *(-300)*	1500 *(-900)*
2017	1800	1400 *(-400)*	800 *(-1000)*
2018	9900	8600 *(-1300)*	6800 *(-3100)*
2019	7700	6800 *(-900)*	5300 *(-2400)*
2020	4800	4100 *(-700)*	2900 *(-1900)*
2021	2100	1800 *(-300)*	1300 *(-800)*
2022	5500	4800 *(-700)*	3400 *(-2100)*
2023	3900	3400 *(-500)*	2700 *(-1200)*

The models were trained using the average temperature data from Helmholtz Munich aggregated on district level with the average of the 10 lowest weeks in the previous year as mortality baseline. The reported results are averages derived from 20 ensemble members. The correction of day-of-the-week was applied. Due to COVID-19, a reliable baseline for mortality rates could not be calculated for 2021-2023; instead, the baseline mortality of the last week in 2020 was used.

Supplementary 4 Contribution of heatwaves to total heat-related mortality

Extended Data Table 8: Comparison of heat-related mortality of the whole year and top 10 days

	whole year	top 10 days	percentage	dates
2014	2000	1500	76%	09-12 Jun, 18-23 Jul
2015	7800	4600	59%	03-07 Jul, 07-10 Aug,14 Aug
2016	2400	1500	62%	24-25 Jun, 20-22 Jul, 26-29 Aug, 14 Sep
2017	1800	1000	54%	29-30 May, 21-24 Jun, 08-09 Jul, 02-03 Aug
2018	9900	5600	57%	27-28 Jul, 31 Jul, 01-05 Aug, 07-08 Aug
2019	7700	4700	60%	26-27 Jun, 30 Jun, 24-29 Jul, 28 Aug
2020	4800	3400	72%	08-15 Aug, 21-22 Aug
2021	2100	1800	88%	17-22 Aug, 28-29 Jun, 14-15 Aug
2022	5500	2400	44%	19 Jun, 20-23 Jul, 04-05 Aug, 14-15 Aug, 17 Aug
2023	3900	2000	50%	22 Jun, 09-12 Jul, 19-23 Aug

The models used were the same as those presented in Extended Data Table 1. Due to COVID-19, a reliable baseline for mortality rates couldn't be calculated for 2021-2023; instead, the baseline mortality of the last week in 2020 was used. We identified the top 10 days with the highest estimated heat-related deaths. These days may not necessarily be the hottest 10 days due to the lag effect in the temperature-mortality response curve.

References

[1] Karine Laaidi, Abdelkrim Zeghnoun, Bénédicte Dousset, Philippe Bretin, Stéphanie Vandentorren, Emmanuel Giraudet, and Pascal Beaudeau. The impact of heat islands on mortality in paris during the august 2003 heat wave. *Environmental health perspectives*, 120(2):254–259, 2012.

[2] Vladimir M Shkolnikov, Ilya Klimkin, Martin McKee, Dmitri A Jdanov, Ainhoa Alustiza-Galarza, László Németh, Sergey A Timonin, Marília R Nepomuceno, Evgeny M Andreev, and David A Leon. What should be the baseline when calculating excess mortality? new approaches suggest that we have underestimated the impact of the covid-19 pandemic and previous winter peaks. *SSM-population health*, 18:101118, 2022.