Article title: Modeling and projecting health-relevant combined ozone and temperature events in present and future Central European climate
Journal name: Air Quality, Atmosphere & Health
Author names: Sally Jahn*, Elke Hertig

***Corresponding author**
E-mail address: **sally.jahn@geo.uni-augsburg.de**
Regional Climate Change and Health, Institute of Geography and Faculty of Medicine, Augsburg University, Alter Postweg 118, 86159 Augsburg, Germany

Table S3 Evaluation metrics used for LR model performance evaluation. Mean and median metric scores across all stations are shown for Central Europe. Minimum and maximum values are given in brackets. The following abbreviations are used: tp for true positives, tn for true negatives, fp for false positives and fn for false negatives.

Metrics	Description	Formula	Mean	Median
True Positive Rate (TPR) *Or* Sensitivity (SN) *Or* Recall (R)	Metric is used to measure the fraction of actual, observed events that are correctly predicted.	$\frac{tp}{tp + fn}$	0.83 [0.74, 0.89]	0.84
True Negative Rate (TNR) *Or* Specificity (SP)	Metric is used to measure the fraction of actual, observed non-events that are correctly predicted.	$\frac{tn}{tn + fp}$	0.81 [0.75, 0.85]	0.81
False Positive Rate (FPR)	Metric is used to measure the fraction of actual, observed non-events thare are incorrectly predicted.	$\frac{fp}{fp + tn}$	0.19 [0.15, 0.25]	0.19
False Negative Rate (FNR)	Metric is used to measure the fraction of actual, observed events thare are incorrectly predicted.	$\frac{fn}{fn + tp}$	0.17 [0.11, 0.26]	0.16
False Alarm Ratio (FARatio)	Metric is used to measure the fraction of the predicted events that did actually not occur.	$\frac{fp}{tp + fp}$	0.50 [0.37, 0.64]	0.49
Precision (P)	Metric is used to measure the fraction of the predicted events that did actually occur.	$\frac{tp}{tp + fp}$	0.51 [0.36, 0.63]	0.51
F1-Score	Metric represents the harmonic mean and thus maintains a balance between recall and precision.	$2 \times \frac{P \times R}{P + R}$	0.63 [0.49, 0.74]	0.63
Area under the ROC curve (AUCROC)	ROC curves show how the number of correctly predicted events varies with the number of incorrectly predicted non-events. AUC measures the entire two-dimensional area underneath curve. Thus, AUC can be used as a summary of the overall model skill.	Area under ROC (Receiver Operating Characteristics) curve created by plotting TPR (y-axis) against FPR (x-axis) at various threshold settings	0.90 [0.85, 0.94]	0.90
Area under the PR curve (AUCPR)	PR curves are similar to ROC curves, but are based on P and R. As both metris are focused on the positive class (events), the area under the PR curve is better to assess the perfomance of the model to predict the minority class. Thus, the metric is especially suitably for highly imbalanced datasets.	Area under PR (Precision-Recall) curve created by plotting P (y-axis) against R (x-axis) at various threshold settings	0.69 [0.5, 0.83]	0.69

Table S4 Changes [%] regarding the amount of days with combined ozone-temperature events between the periods 2031-2050 as well as 2081-2100 compared with 1993-2012 under RCP4.5 and RCP8.5 scenario. Numbers refer to the mean across all stations with respect to each specific ESM. Minimum and maximum values are given in brackets.

	ACCESS1-0	**CanESM2**	**CMCC-CMS**	**CNRM-CM5**	**IPSL-CM5A-MR**	**MIROC-ESM**	**MPI-ESM-MR**
RCP4.5							
2031-2050	17.59 [0.14, 55.00]	12.57 [0.03, 27.21]	4.78 [-5.15, 24.35]	4.93 [-2.59, 15.25]	8.18 [-3.09, 35.13]	7.98 [-9.07, 29.07]	6.52 [-4.54, 33.69]
2081-2100	33.92 [0.25, 95.19]	12.10 [-1.08, 27.11]	6.11 [-5.05, 44.86]	15.13 [0.03, 36.81]	18.38 [-2.58, 56.06]	24.27 [-1.19, 57.84]	7.97 [-9.47, 53.07]
RCP8.5							
2031-2050	29.94 [-1.45,80.00]	14.51 [0.08, 30.73]	-2.13 [-12.44, 13.60]	5.92 [-0.90, 14.20]	14.11 [-0.12, 43.84]	23.74 [0.11, 56.71]	7.25 [-2.06, 30.36]
2081-2100	61.56 [-1.96, 184.10]	39.30 [-8.61, 85.02]	19.01 [-10.47, 93.57]	20.28 [-0.03, 47.92]	51.98 [0.27, 143.62]	49.75 [0.11, 129.49]	20.75 [-9.92, 108.40]