We thank referee #1 for their review and the various suggestions to improve the manuscript. In particular, we appreciate the thorough but constructive critique of our methodology and how it ties into our storyline. In the following we will respond in to the different comments and explain the general changes we intend to make to the manuscript based on them. The reviewer's comments are in black italics, our responses are shown in blue. All line numbers and references refer to the originally submitted manuscript.

The study at hand seeks to attribute the mortality impacts over the UK and the Nordics of the sudden stratospheric warming event in February 2018. To do so, the authors combine state-of-the-art methods from weather prediction modelling with an epidemiological impact analysis. The approach of attributing impacts of single weather events is novel and might help preparing societies better for these consequences. I further want to congratulate the authors on a very clearly written paper that is well structure and thus easy to follow.

We thank the referee for this positive and encouraging comment.

My major concerns are about the epidemiological analysis and its application, the use of monthly data for estimating temperature-mortality relationships, and the application of the different temperature datasets. My review focus is in the health impact analysis, as I'm no expert on stratospheric processes. I've also included some minor comments, that caught my eye whilst reading the draft.

Epidemiological analysis I

To my understanding the authors first fit a model that expresses the expected mortality as a function of (daily mean, I assume) temperature. However, in environmental epidemiology, temperature attributable mortality is calculated differently (see Gasparrini, 2014). Attributable fraction of mortality is calculated as

$AF = (RR(T) - 1) / RR(T)$

where AF stands for attributable fraction and the RR for relative risk from the DLNMs. Temperature attributable numbers of mortality are then subsequently calculated as

$AN = n * AF$

Where n denotes the number of cases (daily deaths, on that day). To be more precise, you could also use the (forward looking) running mean of n, with the number of lag days (same number as lags within the DLNM) as duration.

Directly estimating cold-attributable mortality from the relative risk, as I suspect the authors are doing here, would thus lead to a strong overestimation and is technically incorrect.

We appreciate this important clarification regarding standard epidemiological practice. We agree that, in classical DLNM-based attribution studies using observed daily

mortality, temperature-attributable deaths are typically estimated via attributable fractions derived from relative risks (e.g. Gasparrini, 2014).

Our analysis follows a different objective and interpretation. Rather than estimating attributable fractions like AF from observed daily mortality counts n, we directly compare two counterfactual model simulations (nudged versus control) and translate their temperature differences into expected mortality differences using the DLNM-derived relative risk curve. In this framework, we do not multiply attributable fractions by observed daily deaths. Instead, the relative risk function is used to map simulated temperature anomalies to expected mortality relative to an MMT-based baseline, consistently within each ensemble member.

The resulting estimates therefore represent model-based expected excess deaths associated with the presence of the SSW in a counterfactual framework, rather than epidemiological attributable mortality in the classical sense. For this reason, the standard attributable fraction approach based on observed daily counts is not directly applicable to our study design.

We will revise the manuscript text to clarify this distinction explicitly, to avoid ambiguity and to make clear how our approach differs from standard attributable fraction-based analyses.

Also note that our mortality estimates for the UK are of similar magnitude to those reported by Charlton-Perez et al. (2021) for SSW-related mortality impacts, as referenced in Section 5. This consistency provides additional reassurance that our framework does not lead to a systematic overestimation of effects.

Epidemiological analysis II

Several recent studies use a two-stage approach (i.e. Sera, 2022) two improve the estimate of the coefficient by pooling the spline parameters of the DLNM. This would likely improve your counterintuitive results of the RR curves for some locations as displayed in Figure 7 (i.e. where the curves bend downward again at very cold temperatures). There are several studies with open code which you can use for mimicking the methods (i.e. Vicedo-Cabrera, 2021).

We thank the referee for pointing out recent developments in multivariate meta-analytic DLNM approaches. In some cases, particularly for individual Norwegian cities shown in Fig. S3, the DLNM-derived exposure-response curves do exhibit a downward bending at very cold temperatures. These features occur in temperature ranges with extremely sparse data coverage, where daily death counts are very low and the exposure-response relationship is therefore weakly constrained. From a statistical perspective, such behaviour is expected under spline-based models when extrapolating beyond well-sampled regions of the predictor space and does not necessarily indicate a robust epidemiological signal.

Multivariate meta-analytic DLNM approaches that pool spline parameters across locations (as in Sera, 2022) can indeed stabilise estimates in these data-sparse regimes by borrowing strength from better-sampled regions. However, implementing such pooled models would substantially increase methodological complexity and would shift the focus of the study toward epidemiological model optimisation. Given that our primary aim is event-based attribution rather than methods development, and that we already assess robustness by comparing fundamentally different modeling approaches (DLNM and binning), which yield consistent mortality impact estimates, we do not expect the local behaviour of exposure-response curves at very cold extremes to materially affect our attribution results. We will clarify this interpretation in the revised manuscript, to better link non-monotonic behaviour at very cold temperatures to data sparsity and statistical uncertainty rather than to a robust epidemiological signal.

Epidemiological analysis III - DLNMs for weekly data

The authors state that applying DLNMs to weekly data is not trivial. However, this is routinely being done (i.e. Ballerster et al. (2023)).

Comments III-V are closely related (temporal aggregation, choice of epidemiological model and practical data constraints). We respond to each point below, with some overlap to avoid repetition.

We agree with the referee on the fact that DLNM-type models have been applied to aggregated outcomes such as weekly mortality in the existing literature. We have revised the manuscript to clarify our wording: our point was not that weekly DLNMs are impossible, but that applying DLNMs to aggregated data often requires additional methodological choices (e.g. redefining lag structure and basis specification) that are not central to the aims of the present event-attribution study. We therefore compute DLNMs where daily data are available and use the binning-based approach to assess robustness for coarser temporal aggregation. We have will revise the manuscript text accordingly.

Epidemiological analysis IV – binning methods

While I acknowledge that there is an advantage of using several statistical approaches to estimate temperature-related mortality – I'd suggest restricting the analysis to the standard DLNMs here, as this should not be an epidemiological methods development study. Also, I don't think that it is appropriate to use monthly data for estimating temperature-mortality relationships, as the short term effect of individual days are important. Otherwise, I'd suggest to quantitatively compare models based on their AIC score, and not qualitatively as currently done.

We recognise the referee's preference for DLNMs as the epidemiological standard when daily data are available and we use DLNMs in precisely those settings. The motivation for including the binning approach is to enable a transparent and transferable impact

mapping when only aggregated mortality data are available and/or when implementing an aggregated-outcome DLNM is beyond scope. Importantly, the binning method is used primarily as a robustness check and yields consistent event-attribution estimates with the DLNM where both can be applied, including the UK regions and Norway.

Regarding monthly data: we agree that higher temporal resolution is generally preferable. However, SSW-related circulation anomalies can persist for several weeks to months (see Baldwin and Dunkerton, 2001 or Domeisen et al., 2020 cited in the manuscript), which makes it more plausible that aggregated mortality products retain part of the relevant signal in this specific application. We will revise manuscript text to clarify this rationale and to emphasise that daily/weekly data are preferred when accessible.

Epidemiological analysis V – country level and monthly data

Again, I'd suggest to sticking to standard epidemiological methods, as otherwise you should show a clear model intercomparison (local vs. district/country vs. country level results, daily vs. weekly vs. monthly data, model assumption, model parameterization, etc.). The argument of using country level and monthly data for Nordic countries is not very convincing to me, as there certainly are daily mortality counts somewhere (at least there is published literature on it).

The primary objective of this paper is event-based attribution across multiple regions and data environments. While higher-resolution mortality data may exist for parts of the Nordic countries, such series are not always publicly available, harmonised in a way that allows consistent multi-country analyses over long periods. We therefore use the highest temporal resolution available to us for each region and complement this with aggregated datasets to test sensitivity to temporal resolution.

We will further clarify this data-availability and harmonisation rationale in manuscript. As part of this, we will further emphasise in the revised text that higher temporal resolution data is preferred where available and that we interpret monthly-resolution results primarily as a robustness/sensitivity assessment rather than as a replacement for standard daily DLNM analyses.

Temperature data applied

I see two concerns with the temperature data – I don't think it's adequate to use a country level mean as input to the epidemiological analysis, especially for Finland, where a vast majority of the population is concentrated around Helsinki. I'd suggest using a population weighted temperature time series. Same could be done on a county-level. But then again, I don't think that a country wide assessment is suitable in the first place.

We agree that population-weighted temperature series can, in principle, better represent individual exposure. However, in the present case the SSW-induced temperature anomalies are spatially large-scale and highly coherent across each country (Figs. 1 and 2), with only weak spatial gradients relative to the scale of population density variations. We therefore expect country-mean temperatures to provide a reasonable approximation of the exposure relevant for country-level mortality impacts in this specific event-based framework.

Constructing population-weighted temperature series would additionally require combining model output and gridded population datasets on a common grid (e.g. via spatial interpolation of both fields), which introduces further methodological choices without substantially altering the large-scale anomaly pattern that drives our attribution results.

We will clarify this reasoning the text of the revised manuscript.

The spatial resolution column in Table 3 is not very helpful (for me at least). Could you maybe express everything in degrees (as being done for the CESM2)?

We will add approximate degree-representations of the spatial resolution to Table 3.

It is quite important that the temperature distribution of the model data corresponds to the distribution with which the temperature-mortality relationships are calculated, otherwise you might include a model-bias in your impact calculation. This can be done i.e. using quantile mapping as a bias-correction method.

Regarding bias correction, mortality impacts are inferred from differences between nudged and control simulations, such that systematic biases in absolute temperature largely cancel in the attribution step. We acknowledge that, due to the nonlinear nature of temperature–mortality relationships, this cancellation is not expected to be mathematically exact. However, as both experiments share the same model physics and baseline temperature distribution, the remaining bias in the differential signal is expected to be small relative to the SSW-induced anomaly.

The robustness of this approach is further supported by the consistency of results across multiple independent forecast systems. Moreover, the SNAPSI experiments do not provide a long-term model climatology, which precludes the application of standard bias-correction methods such as quantile mapping. We will revise the manuscript text to clarify this point.

Minor comments

Section 3.3 is a bit of a blend of results and discussion, yet is listed within the Method Section

We agree with the referee that Section 3.3 contains descriptive results and interpretative elements. We will revise the manuscript text and structure.

Page 11 – typo: "..exposure-response curve with strongly depend.." (with instead of will)

We will fix this.

Page 13: MMM (I assume mulit model mean) – I think it's the first time that this acronym appears within this manuscript

We will add the full term.

References:

Charlton-Perez, A.J., Huang, W.T.K. and Lee, S.H., 2021. Impact of sudden stratospheric warmings on United Kingdom mortality. Atmospheric Science Letters, 22(2), p.e1013.

Gasparrini, A., & Leone, M. (2014). Attributable risk from distributed lag models. BMC medical research methodology, 14(1), 55.

Sera, F., & Gasparrini, A. (2022). Extended two-stage designs for environmental research. Environmental health, 21(1), 41.

Ballester, J., Quijal-Zamorano, M., Méndez Turrubiates, R.F. et al. Heat-related mortality in Europe during the summer of 2022. Nat Med 29, 1857–1866 (2023). https://doi.org/10.1038/s41591-023-02419-z