Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Stratospheric events and long-range Scandinavian winter surface temperature forecasts

Note no SAMBA/21/19

Authors
Kristoffer Skuland
Claudio Heinrich
Alex Lenkoski
Thordis L. Thorarinsdottir

Date 9th August 2019

Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

The authors

Kristoffer Skuland is a master's student in industrial mathematics at the Norwegian University of Science and Technology, Claudio Heinrich is Research Scientist at Norwegian Computing Center, Alex Lenkoski is Chief Research Scientist at Norwegian Computing Center and Thordis L. Thorarinsdottir is Chief Research Scientist at Norwegian Computing Center.

Norwegian Computing Center

Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Norsk Regnesentral
Norwegian Computing Center
Postboks 114, Blindern
NO-0314 Oslo, Norway

Besøksadresse
Office address
Gaustadalléen 23
NO-0373 Oslo, Norway

Telefon · telephone
(+47) 22 85 25 00
Telefaks · telefax
(+47) 22 69 76 60

Internett · internet
www.nr.no
E-post · e-mail
nr@nr.no

Title	**Stratospheric events and long-range Scandinavian winter surface temperature forecasts**
Authors	**Kristoffer Skuland** <kristoffer.skuland@gmail.com> **Claudio Heinrich** <claudio.heinrich@nr.no> **Alex Lenkoski** <alex@nr.no> **Thordis L. Thorarinsdottir** <thordis@nr.no>
Date	9th August 2019
Publication number	SAMBA/21/19

Abstract

At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio.

Keywords	seasonal forecast, SSW, SPV, NAO, stratosphere
Target group	Climate scientists, users of climate information
Availability	Open
Project	Seasonal forecasting engine
Project number	220857
Research field	Statistics, climate science
Number of pages	33

Contents

1 Introduction

Complex weather systems and vigorous circulation in the troposphere makes predicting the weather more than one week in advance a difficult topic, and this problem is especially severe during the winter. The stratosphere is however far more stable, and might have some effect on surface weather patterns: Strong variations in the straospheric circulation can descend into the troposphere and cause La Niña and El Niño-like climate patterns over the North Atlantic and Scandinavia (see e.g. Baldwin and Dunkerton (2001)).

During winters strong stratospheric westerly winds, called the stratospheric polar vortex, circulate around cold air over the Arctic. The polar vortices strengthen and weaken from year to year, and in some years the winds in the polar vortex temporarily weakens sufficiently to that effect that the vortex reverse to flow from east to west; the flow of Arctic air becomes more disorganized, and masses of cold air sinks from the stratosphere into the troposphere, causing rapid and sharp temperature drops. The breakdown of the polar vortex is an extreme event known as a *sudden stratospheric warming* (SSW), and the converse, sufficiently strong polar vortices, is an extreme event known as a *strong polar vortex* (SPV).

Previous studies have indicated significant forecast skill of the timing of sudden stratospheric warmings to around two weeks ahead, but Scaife et al. (2016) indicates that there exits a predictability of SSWs and SPVs beyond the deterministic range. Identifying how and to what degree stratospheric events affect surface temperatures, this could then be used to improve medium and long-range temperature forecasts.

This note starts by reproducing the results of Scaife et al. (2016) section 2, before we extend this analysis to include a more rigorous inquiry into the predictive skill of stratospheric polar vortex forecasts from *The Met Office* (UKMO) and *The European Centre for Medium-Range Weather Forecasts* (ECMWF). We then go on to study how extreme stratospheric events affect Scandinavian surface temperatures in December–March by creating a non-parametric impulse function, before we attempt to improve medium and long-range surface temperature forecasts by taking into account long-range forecasts of the stratospheric polar vortex from UKMO and ECMWF.

2 Data

2.1 Ensemble forecast systems

Ensemble forecasts of stratospheric westerly winds were aquired from *The Met Office* (UKMO), the United Kingdom's national weather service, and *The European Centre for Medium-Range Weather Forecasts* (ECMWF), where the UKMO forecast had 7 ensemble members and the ECMWF forecast had 24 ensemble members.

2.2 Observational data

Zonal mean U wind and temperature observational data from ...

2.3 Preparation of data

The forecast data from UKMO and ECMWF are stored in different GRIB-files and have to be processed before they can be analysed. All wanted zonal wind data for 60° longitudes and all latitudes are extracted, while all unwanted data is removed. Only forecasts from 1st November 00:00 to 2nd March 00:00 for each winter from 1993/94 to 2015/16 are saved, with one new forecast every 12th hour for each ensemble member. The zonal wind forecasts in the stratosphere is averaged over all latitudes for each forecast at a given lead time. For each forecast system this gives a $4209 \times K$ matrix, for K ensemble members.

High-resolution shapefiles of Norway, Sweden, Denmark and Finland's administrative county and municipal divisions were downloaded from the *Database of Global Administrative Areas* (gadm.org), and the shapefiles were followingly simplified (mapshaper.org) and exported to geoJSON-files which were used to extract temperature measurements over mainland Scandivania from a .nc-file containing daily temperature measurements in every 1/4 longitude and latitude. For each lead time, the temperature was averaged over all reported observations within the administrative division.

3 Theory

3.1 Scoring rules

We denote a mean zonal wind observation by $y \in \Omega$, where $\Omega = \mathbb{R}$. Similarily, for observations of SSW and SPV events, y is a random boolean observation with $\Omega = \{0, 1\}$. A probabilistic prediction for y is given by a distribution function with support on Ω denoted by $F \in \mathcal{F}$ for some appropriate class of distributions $\mathcal{F}$.

Probabilistic forecast accuray is normally assesed using *scoring rules*, which assign a numerical score to each forecast–observation pair, where a lower penalty indicates better predictive preformance. Specifically, a scoring rule is a mapping

$$S : \mathcal{F} \times \Omega^d \to \mathbb{R} \cup \{\infty\}. \tag{1}$$

A scoring rule is *proper* relative to class $\mathcal{F}$ if

$$\mathbb{E}_G S(G, Y) \leq \mathbb{E}_G S(F, Y) \tag{2}$$

for all probability distributions $F, G \in \mathcal{F}$. That is, the expected score for a random observation Y is optimized if the true distribution of Y is issued as a forecast.

3.2 Brier scores

The *Brier score* is a proper scoring rule which assesses the predictive probability of treshold exceedance. The Brier score is usually written in the form

$$\text{BS}(F, y|u) = (p_u - \mathbb{1}\{y \geq u\})^2 \tag{3}$$

for a treshold u with $p_u = 1 - F(u)$.

3.3 CRPS

The *continuous ranked probability score* (CRPS) is of particular interest in that it simultaneously assesses both calibration and sharpness, and thus all three types of goodness discussed by Murphy (1993). The CRPS applies to probability distributions with a finite mean and is defined by

$$\text{CRPS}(\widehat{F}, t) = \mathbb{E}_{\widehat{F}}|X - t| - \frac{1}{2}\mathbb{E}_{\widehat{F}}\mathbb{E}_{\widehat{F}}|X - X'|, \tag{4}$$

where $\hat{F}$ is a forecast distribution with a finite first moment and $X, X' \sim \widehat{F}$ denote two independent random variables. For an ensemble $\mathbf{x} = \{x_1, \ldots, x_K\}$, the CRPS equals

$$\text{CRPS}(\mathbf{x}, t) = \frac{1}{K}\sum_{k=1}^{K}|x_k - t| - \frac{1}{2K^2}\sum_{k=1}^{K}\sum_{l=1}^{K}|x_k - x_l|. \tag{5}$$

3.4 MSE

A similar and simpler scoring rule assessing the marginal accuracy is the mean squared error (MSE),

$$\mathrm{MSE}(\widehat{F}, t) = (\widehat{\mu} - t)^2, \tag{6}$$

where $\widehat{\mu}$ is the mean of $\widehat{F}$.

3.5 Uncertainty and significance of skill scores

The estimation of the mean score may be associated with a large uncertainty, and a bootstrapping procedure over the individual scores may be utilised in order to assess the uncertainty in the mean score (Friederichs and Thorarinsdottir, 2012). Assume we have n score values $S(F_1, y_1), \ldots, S(F_n, y_n)$. By repeatedly resampling vectors of length n with replacement and estimating the mean of each sample, we obtain an estimate of the variability in the mean score.

Similarily, to test the significance of score differences between two competing methods, we can apply a permutation test relying on resampling (Good, 2013; Möller et al., 2013). Two competing predictive distributions $\widehat{F}_1$ and $\widehat{F}_2$ are compared under a scoring rule $S(F, \cdot)$ using the statistic

$$s := \frac{1}{n} \sum_{i=1}^{n} \left(S(\widehat{F}_1, y_i) - S(\widehat{F}_2, y_i) \right). \tag{7}$$

The permutation test is then based on resampling copies of s with the labels of $\widehat{F}_1$ and $\widehat{F}_2$ swapped for a random number of summands.

3.6 Kaplan-Meier estimator

The Kaplan-Meier estimator for the survival function $S(t) = P(T > t)$, the probability that the time of an event is later than some specified time t, is

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i} \right), \tag{8}$$

for time t_i when d_i events occured, and n_i is the number of individuals at risk just before t_i. The variance of the Kaplan-Meier estimator is estimated by Greenwood's formula,

$$\widehat{\mathrm{Var}}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \tag{9}$$

The difference between two survival curves estimated by Kaplan-Meier, denoted by $i = \{1, 2\}$ with events occuring and number of individuals known to have survived up to time j being d_{ij} and n_{ij}, may be tested by a logrank test with test statistic

$$Z = \frac{\sum_{j=1}^{J}(d_{ij} - E_{ij})}{\sqrt{\sum_{j=1}^{J} V_{ij}}} \xrightarrow{d} N(0,1) \tag{10}$$

where at least one event occurs at each time step $j = 1, \ldots, J$, and $E_{ij} = d_j n_{ij}/n_j$ and $V_{ij} = E_{ij}(1 - n_{ij}/n_j)((n_j - d_j)/(n_j - 1))$ for $n_j = n_{1j} + n_{2j}$ and $d_j = d_{1j} + d_{2j}$.

4 Results and discussion

4.1 Reproducing Scaife et al. (2016) section 2

To quantify the predictive skill of SSW and SPV events, Scaife et al. (2016) compute what they call the *perfect predictability* of the UKMO ensemble forecast, which attempts to estimate the predictability of events using only the forecast data alone[1]. We firstly reproduce these results using three more years of data and an extra forecast system[2].

Similar to Scaife et al. (2016) we define a SSW to occur when the zonal mean U wind (the daily zonal Artic winds at 10 hPa, 60 °N and averaged over all longitudes) decreases below zero, while SPV events are defined to occur if the zonal mean U wind increases above 48 ms^{-1}.

For $b = 1, \ldots, B$ bootstrap samples and $y = 1, \ldots, Y$ years, a proxy observation $P_y^{(b)}$ is drawn randomly from the K ensemble members, $P_y^{(b)} \in \{E_{y1}, \ldots, E_{yK}\}$. The increased risk of a SSW (or equivalently a SPV) event in the years when an event occurred in the proxy observations, is given by the difference between the two probabilities $P(SSW \in E_k | SSW \in P)$ and $P(SSW \in E_k | SSW \notin P)$. For an out-of-sample computation the first probability is estimated by, using Bayes' rule,

$$\begin{aligned} P(SSW \in E_k | SSW \in P) &= \frac{P(SSW \in E_k, SSW \in P)}{P(SSW \in P)} \\ &= \frac{\frac{1}{BY(K-1)} \sum_{b=1}^{B} \sum_{y=1}^{Y} \sum_{k:E_{ky} \neq P_y^{(b)}} I(SSW \in E_{yk}, SSW \in P_y^{(b)})}{\frac{1}{BY} \sum_{b=1}^{B} \sum_{y=1}^{Y} I(SSW \in P_y^{(b)})}. \end{aligned} \tag{11}$$

For an in-sample computation the third sum in the numerator is over all k, and the numerator is divided by K instead of $K - 1$.

The estimated probabilities for both out-of- and in-sample perfect predictability for ECMWF, UKMO and the combined UKMO/ECMWF ensemble forecast are presented in table 1. We observe that the out-of-sample UKMO probabilities likens those reported by Scaife et al. (2016), who observed a 12 % rise in the forecast probability of an event on average from 47 % in winters in which no event occured to 59 % in winters in which an event occured. Since Scaife's perfect probabilities in part are clumsy ways to communicate the same information as Brier scores, the forecast systems' Brier scores are for completeness too included table 1. These Brier scores are however not comparable between ensemble forecasts, since they are based on different sets of proxy outcomes.

1. This peculiar approach is rooted in a "metrological paradigm", which treats (numerical) weather forecast not as a statistical problem, but assumes away all shortcomings of the model and treats the forecast as an initial value problem. Scaife et al. (2016) is consequently not that interested in testing whether or not the ensemble forecasts have actual predictive skill, but if different realisations from the same numerical model, with similar initial values, are correlated.
2. Furthermore, Scaife et al. (2016) analysed an ensamble forecast from UKMO with 24 ensemble members, whereas our ensemble forecast from UKMO only contains 7 ensemble members.

Table 1. Estimates of out-of-sample and in-sample perfect predictability in the ECMWF, UKMO and combined ECMWF/UKMO ensemble forecast systems, as well as non-comparable proxy Brier scores added for completeness.

	ECMWF		UKMO		Combined	
	Out-of	In	Out-of	In	Out-of	In
$P(SSW \in E_k \mid SSW \in P)$	0.71	0.72	0.61	0.67	0.65	0.67
$P(SSW \in E_k \mid SSW \notin P)$	0.66	0.64	0.49	0.42	0.62	0.57
$P(SPV \in E_k \mid SPV \in P)$	0.60	0.61	0.64	0.69	0.63	0.65
$P(SPV \in E_k \mid SPV \notin P)$	0.52	0.50	0.51	0.44	0.52	0.48
$P(SPV \in E_k \mid SSW \in P)$	0.55	0.55	0.57	0.57	0.58	0.57
$P(SPV \in E_k \mid SSW \notin P)$	0.59	0.59	0.60	0.60	0.58	0.59
$\overline{BS}_{SSW}$	0.212	0.195	0.251	0.184	0.233	0.201
$\overline{BS}_{SPV}$	0.236	0.218	0.246	0.181	0.246	0.212

Table 2. Estimates of predictability in the UKMO, ECMWF and combined UKMO/ECMWF forecasts conditioned on observations. Bootstrap estimates of standard deviations of the probabilities in parentheses.

	UKMO	ECMWF	Combined
$P(SSW \in E_k \mid SSW \in Y)$	0.65 (0.06)	0.69 (0.04)	0.68 (0.04)
$P(SSW \in E_k \mid SSW \notin Y)$	0.46 (0.07)	0.69 (0.04)	0.57 (0.06)
$P(SPV \in E_k \mid SPV \in Y)$	0.54 (0.07)	0.59 (0.05)	0.59 (0.05)
$P(SPV \in E_k \mid SPV \notin Y)$	0.55 (0.08)	0.51 (0.05)	0.53 (0.07)
$P(SPV \in E_k \mid SSW \in Y)$	0.52 (0.08)	0.55 (0.06)	0.56 (0.06)
$P(SSW \in E_k \mid SSW \notin Y)$	0.56 (0.07)	0.55 (0.05)	0.56 (0.06)

4.2 Scaife et al. (2016) section 2 with actual predictive assessment

We next perform the same analysis as in section 4.1, but now the probabilities of an event occuring in an ensemble member are conditioned on actual observations and not proxy observations. To compare with the true observations, we only consider the forecasts for 12:00 each day. The climatological frequency of SSW and SPV events in the UKMO, ECMWF and the combined forecast are 0.54 and 0.55, 0.69 and 0.55, and 0.62 and 0.56, respectively, while the actual climatological frequency of SSW and SPV events in the same period of time is 0.48 and 0.52. The estimated probabilities are given in table 2.

Compared to table 1, we now observe that conditioning on actual events instead of proxy observations has diminished all apparent skill in predicting SSWs (and likely too SPVs) in the ECMWF ensemble forecast. UKMO has no skill in predicting SPVs, but appears to retain some skill in predicting SSWs since, for $\mu := P(SSW \in E_k \mid SSW \in Y) - P(SSW \in E_k \mid SSW \notin Y)$, the one-tailed hypothesis test

$$H_0 : \mu = 0 \quad \text{vs.} \quad H_1 : \mu > 0 \tag{12}$$

performed by a permutation test, has p-value 0.033.

Similarily, figure 1 displays the ensemble forecast probabilities for the occurance of at least one SSW event for each year in the UKMO and ECMWF ensemble forecasts, with a colouring indicating whether or not an actual event did occur in the given year. We observe that ECMWF lacks a spread in their forecast probabilities, and again that ECMWF appears to have no significant skill in predicting the occurance of SSWs.

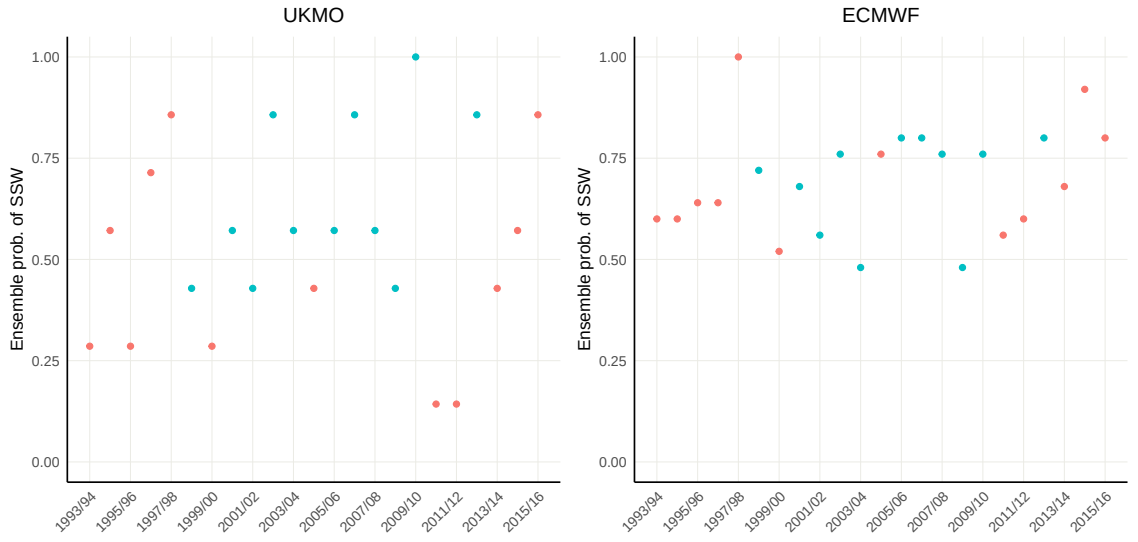

Figure 1. The ensemble forecast probability of a SSW event for each year in the UKMO and ECMWF ensemble forecasts, where a red colour means that no SSW event did occur in the given year, while a blue colour indicates that a SSW event did occur in the given year.

4.3 Accuracy of probabilistic predictions of events

The accuracy of probabilistic predictions of events is estimated by the Brier score, where the probability of at least one event occuring in a forecast system is estimated by the frequency of ensemble members containing at least one event. The mean Brier scores, averaged over all years, for climatology, UKMO, ECMWF and the combined UKMO/ECMWF forecast are given in table 3.

In order to assess whether the mean Brier scores are significantly different from each other, (two-tailed) permutation tests may be applied. These tests, however, show that the mean Brier scores of UKMO, ECMWF and Combined are not significantly different from climatology, see table 3. These results furthermore appear to hold for shorter time intervals.

In the competition between the different ensemble forecasts, UKMO is found to be different from ECMWF in predicting SSWs with p-value 0.048, and UKMO is found to be different from Combined in predicting SSWs with p-value 0.040 and in predicting SPVs with p-value 0.084.

Table 3. Mean Brier scores, averaged over all years, for climatology, UKMO, ECMWF and the combined UKMO/ECMWF ensemble forecast. p-values from hypothesis tests of equality versus non-equality of each Brier score to climatology's Brier score, stands in parenthesis behind each skill score.

	$\overline{BS}_{SSW}$	$\overline{BS}_{SPV}$
Climatology	0.27	0.27
UKMO	0.22 (0.260)	0.32 (0.464)
ECMWF	0.31 (0.514)	0.24 (0.282)
Combined	0.28 (0.980)	0.26 (0.748)

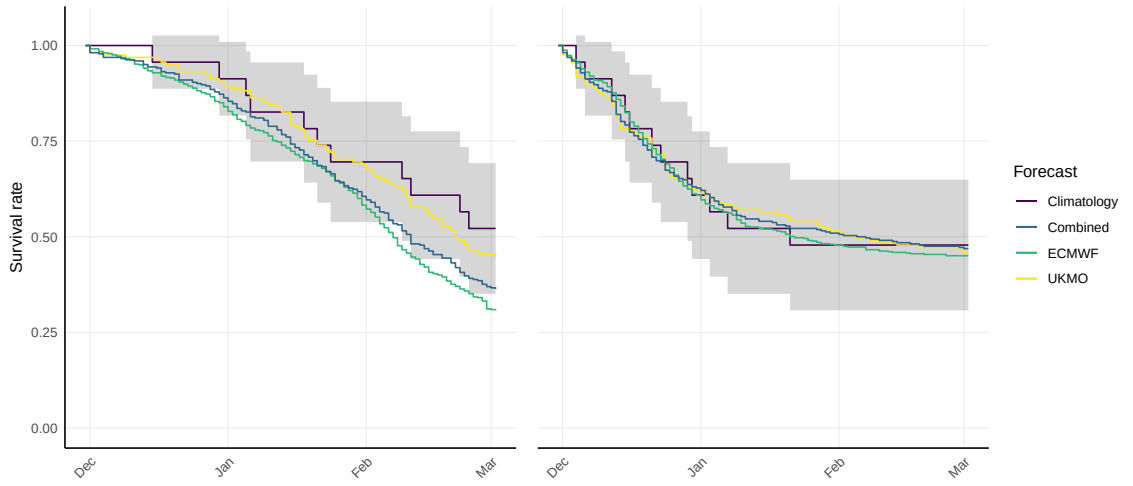

Figure 2. Kaplan-Meier estimators for the survival curves of UKMO, ECMWF and combined UKMO/ECMWF in predicting the occurance of SSWs (left panel) and SPVs (right panel), compared with the historical survival curve (climatology). The grey region is the estimated 90 % confidence region of the climatology survival curve.

4.4 Event survival analysis

We wish to assess if our ensemble forecasters correctly represent the hazard that at least one event has occured by a given day of the winter. Hence, for all years and ensemble members, we count the number of days from 1st December until an event occurs, and then we compute the survival curve of said event, which we estimate by the Kaplan-Meier estimator. The variance of the Kaplan-Meier estimator is estimated by Greenwood's formula.

The estimated survival curves of UKMO, ECMWF, combined UKMO/ECMWF and climatology in predicting SSWs and SPVs are presented in figure 2, and the p-values of logrank tests between the different survival curves and the climatology survival curve is presented in table 4. ECMWF appears to overshoot the hazard of a SSW at every lead time, and its survival curve is (approximately) found to be different from that of climatology under a 5 % significance level. Furthermore, the survival curve of UKMO in predicting SSWs is found to be different than that of ECMWF with p-value 1.1e-3. If we only look at the month of December, however, ECMWF is not significantly different from climatology in predicting SSWs. In predicting SPVs both UKMO and ECMWF appears to comply well with the historic survival curve.

Table 4. p-values from two-tailed logrank tests with null hypothesis that the Kaplan-Meier estimated survival curves in figure 2 of the given forecaster, in predicting either SSWs or SPVs, is equal to that of climatology.

	SSW	SPV
Climatology	–	–
UKMO	0.541	0.887
ECMWF	0.057	0.828
Combined	0.175	0.852

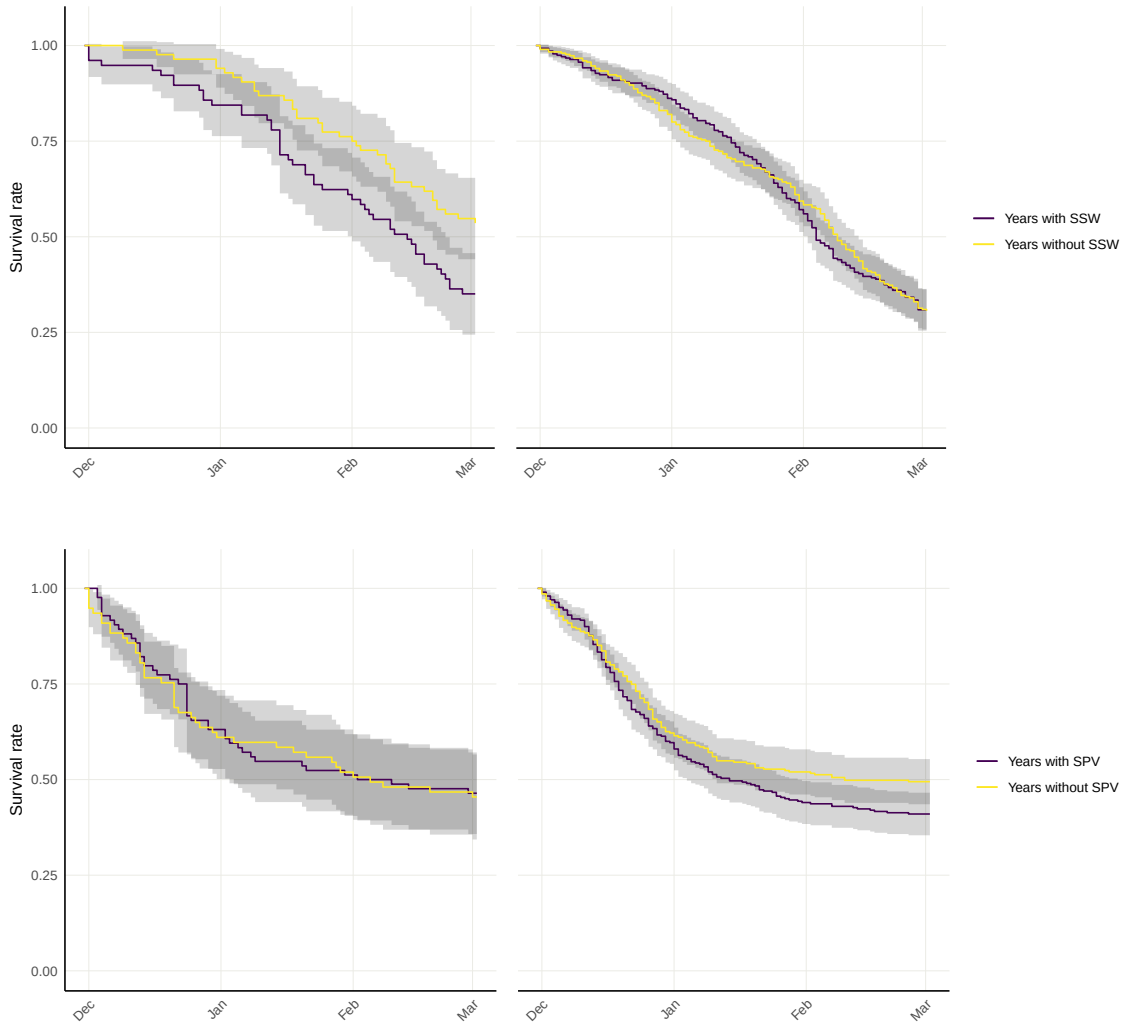

Figure 3. Kaplan-Meier estimators for the survival curves of UKMO (left panel) and ECMWF (right panel) in years when a SSW/SPV did and did not take place. The grey region is the estimated 95 % confidence region of the survival curves.

In figure 3 we have estimated the survival curves for UKMO and ECMWF in years which a SSW occured and in years which no such event occured. Again we observe that ECMWF appears to have little skill in predicting whether or not an event will occur, and the two curves are not significantly different; UKMO however reports significantly different – with p-value 0.011 – survival curves for years with and without reported SSW events.

A similar analysis for SPVs is too presented in figure 3, where ECMWF perform stronger than UKMO January–March, and the two ECMWF curves are found to be different with p-value 0.072. UKMO appears to have no skill in differentiating the event risk for years in which events did and did not occur.

4.5 Predictive skill of zonal mean U wind forecasts

The predictive skill of mean zonal wind forecasts is assessed by mean squared error and continuous ranked probability score. MSEs and CRPS' for daily, weekly and monthly

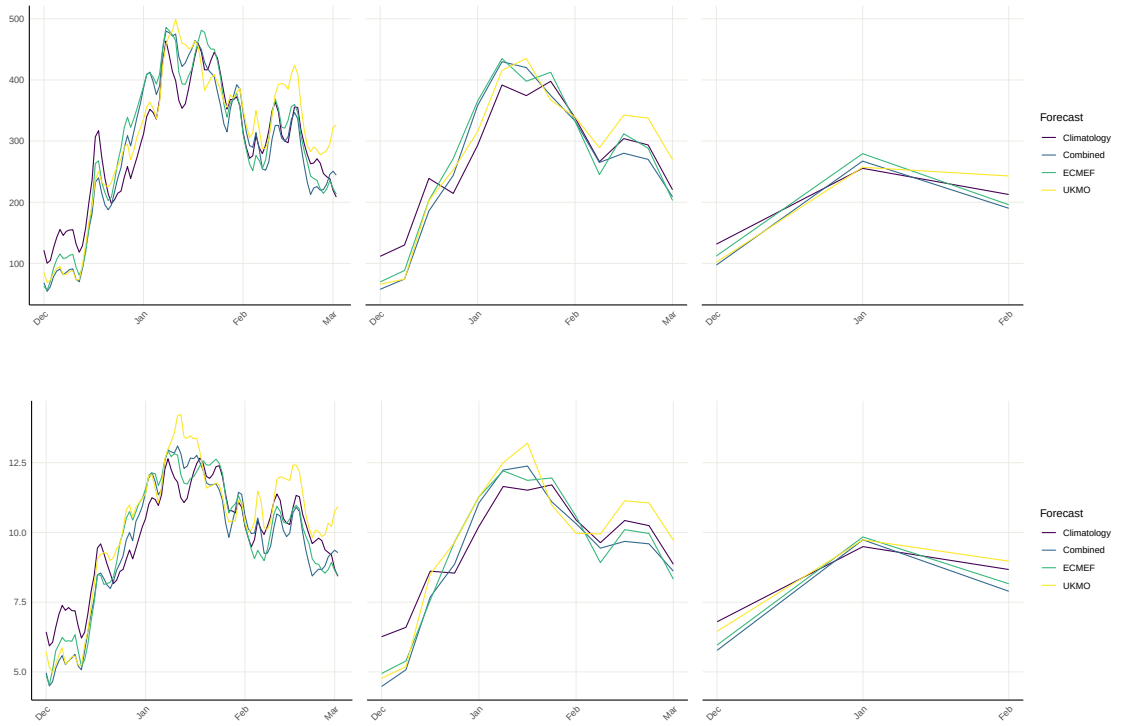

Figure 4. Predictive skill of zonal wind forecasts by climatology, UKMO, ECMWF and the combined UKMO/ECMWF ensemble forecast, for daily, weekly and monthly resolutions, measured in MSE (top row) and CRPS (bottom row), averaged over all winters.

resolutions at each lead time, averaged over all observed years, for climatology, UKMO, ECMWF and the combined UKMO/ECMWF ensemble forecast are presented in figure 4. The MSE and CRPS averaged over all lead times and years at different resolutions for each forecast system, is presented in table 5 and 6.

In order to assess the significance of the differences between the mean MSEs and the mean CRPS', the hypothesis test

$$H_0 : \overline{MSE}_r^i = \overline{MSE}_r^{\text{Climatology}} \quad vs. \quad H_1 : \overline{MSE}_r^i \neq \overline{MSE}_r^{\text{Climatology}}, \tag{13}$$

for $i = \{\text{UKMO}, \text{EMCWF}, \text{Combined}\}$ and resolutions $r = \{\text{daily}, \text{weekly}, \text{monthly}\}$, may be performed using a permutation test. The results of the given hypothesis test is reported in parenthesis behind each skill score in table 5 and 6, and reports that UKMO is significantly different from climatology over the winter at a 1 % significance level.

From figure 4 we, however, observe that UKMO and ECMWF consistently beat climatology during the first two–three weeks of December; a permutation test assessing the difference in skill scores MSE_t and $CRPS_t$, at each lead time t, between UKMO and climatology for daily, weekly and monthly resolutions is presented in figure 5. Performing the aforementioned hypothesis test in equation (13), with MSE and CRPS skill scores averaged over the first three weeks of December for a daily lead time resolution, UKMO,

Table 5. The predictive skill of zonal wind forecasts by climatology, UKMO, ECMWF and combined UKMO/ECMWF ensemble forecast, assessed by mean squared error for daily, weekly and monthly lead time resolution, averaged over all years and lead times. p-values from the (two-tailed) hypothesis test in equation (13) is reported in parenthesis behind each skill score.

	$\overline{MSE}_{\text{daily}}$	$\overline{MSE}_{\text{weekly}}$	$\overline{MSE}_{\text{monthly}}$
Climatology	293.67	274.97	199.87
UKMO	306.53 (0.010)	285.16 (0.379)	200.70 (0.375)
ECMWF	298.19 (0.268)	278.81 (0.699)	195.84 (0.996)
Combined	301.18 (0.000)	280.68 (0.0704)	197.66 (0.383)

Table 6. The predictive skill of zonal wind forecasts by climatology, UKMO, ECMWF and combined UKMO/ECMWF ensemble forecast, assessed by continuous ranked probability score for daily, weekly and monthly lead time resolution, averaged over all years and lead times. p-values from the (two-tailed) hypothesis test in equation (13) is reported in parenthesis behind each skill score.

	$\overline{CRPS}_{\text{daily}}$	$\overline{CRPS}_{\text{weekly}}$	$\overline{CRPS}_{\text{monthly}}$
Climatology	9.91	9.56	8.32
UKMO	10.27 (0.001)	9.84 (0.385)	8.39 (0.510)
ECMWF	9.78 (0.112)	9.44 (0.496)	7.99 (0.374)
Combined	9.93 (0.060)	9.55 (0.701)	8.16 (0.757)

ECMWF and the combined UKMO/ECMWF ensemble forecast are all significantly different from climatology at a 1 % significance level.

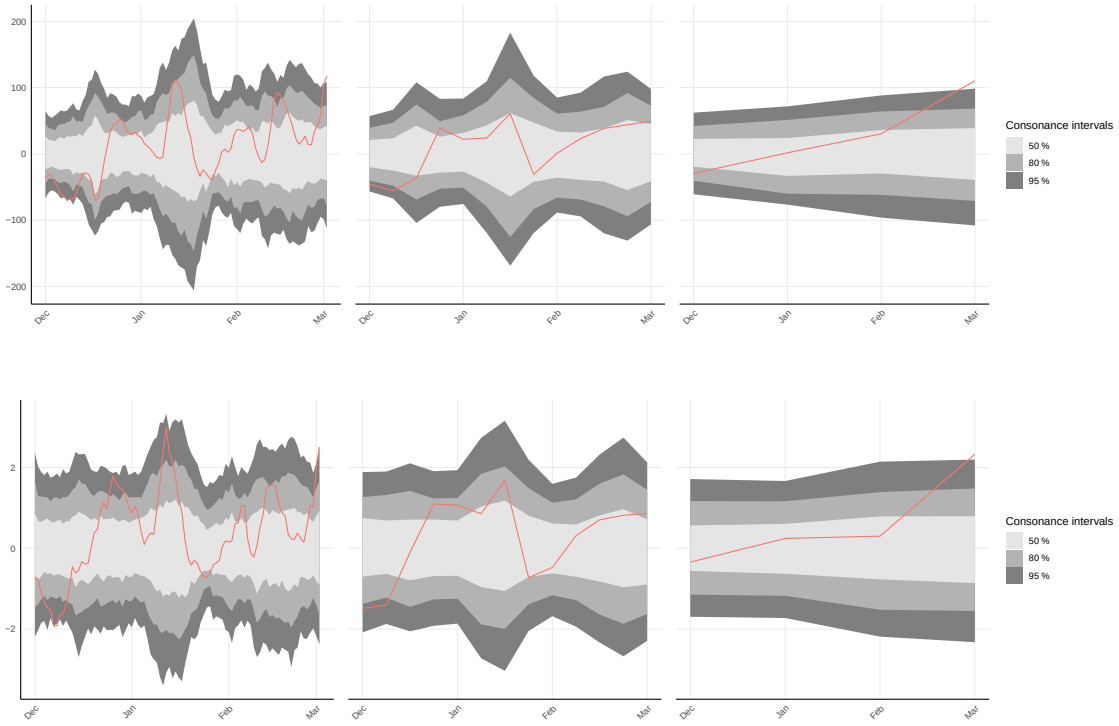

Figure 5. Permutation tests assessing the significance of the difference between the UKMO and climatology forecasts at each lead time, for daily, weekly and monthly resolutions. The top row tests the difference in MSE skill scores, while the bottom row tests the difference in CRPS skill scores. Shaded areas give 50 %, 80 % and 95 % consonance intervals. The red line is the observed difference between the forecasts.

4.6 Impulse function of SSW and SPV events

Having observed the daily mean temperature in each Norwegian, Swedish, Danish and Finnish county from winter 1993–94 to winter 2015–16, we wish to assess the impact a SSW or a SPV event has on the surface temperature n days after the event took place. We thus want to compute the mean temperature anomaly with respect to climatology n days after an event was observed ("the impulse function of SSWs"),

$$I_{r,n} = \frac{1}{|(y,d) \in SSW|} \sum_{(y,d) \in SSW} (t_{r,y,d+n} - \overline{t_{r,y,d+n}}), \quad \forall n, r \tag{14}$$

where $(y,d) \in SSW$ denotes days containing a SSW, $t_{r,y,d+n}$ is the mean temperature in region r at year y and day $d+n$, and $\overline{t_{r,y,d+n}}$ is the climatology for temperature in region r at year y and day $d+n$, given by

$$\overline{t_{r_0,y_0,d_0}} = \frac{1}{Y-1} \sum_{y \neq y_0} t_{r_0,y,d_0}, \tag{15}$$

for a total number of Y years. The impulse functions of SSWs and SPVs in Norwegian, Swedish, Danish anf Finnish counties are presented in figure 6 and 7.

Herein we observe that a SSW is related to a negative temperature anomaly which reaches its maximum after about 10 days and lasts about 30 days after a event in Norway, Sweden and Denmark, while the temperature effect is less pronouced in Finland. Similarily, a SPV is related to a positive temperature anomaly which reaches its maximum after about 10 days and lasts about 20–30 days after the event. In Norway, e.g., the mean temperature anomalies are closely correlated between all counties, but the temperature effect is strongest in Østlandet and weakest along the coast and in Nord-Norge.

4.7 Improving surface temperature forecasts by event climatology

Building on section 4.6, we wish to assess if the impulse function of SSWs and SPVs can be used to improve surface temperature forecasts in Scandinavian counties (Norwegian *fylke*, Swedish *län*, Finnish *maakunta* and Danish *region*) following the observation of an event. That is, immediately following an observation of a SSW or a SPV event, we attempt to forecast the next 30 days based on an *event climatology* equal to the impulse function of said event.

Figure 8 and 9 shows the MSE in Oslo and Finnmark at lead times one to thirty days following the observation of a SPV event. In both counties the event climatology appears to outperform the standard temperature climatology during the first two–three weeks after an event was observed. We can also observe that the form of the MSE curves varies geographically: South of Trøndelag, counties behave similarily to Oslo, where they struggle the most at predicting surface temperature the first days following a SPV; in Nord-Norge the curves however likens that of Finnmark, where the climatologies perform fairly well immediatly following a SPV, before their predictive skill detoriates during the next two–three weeks.

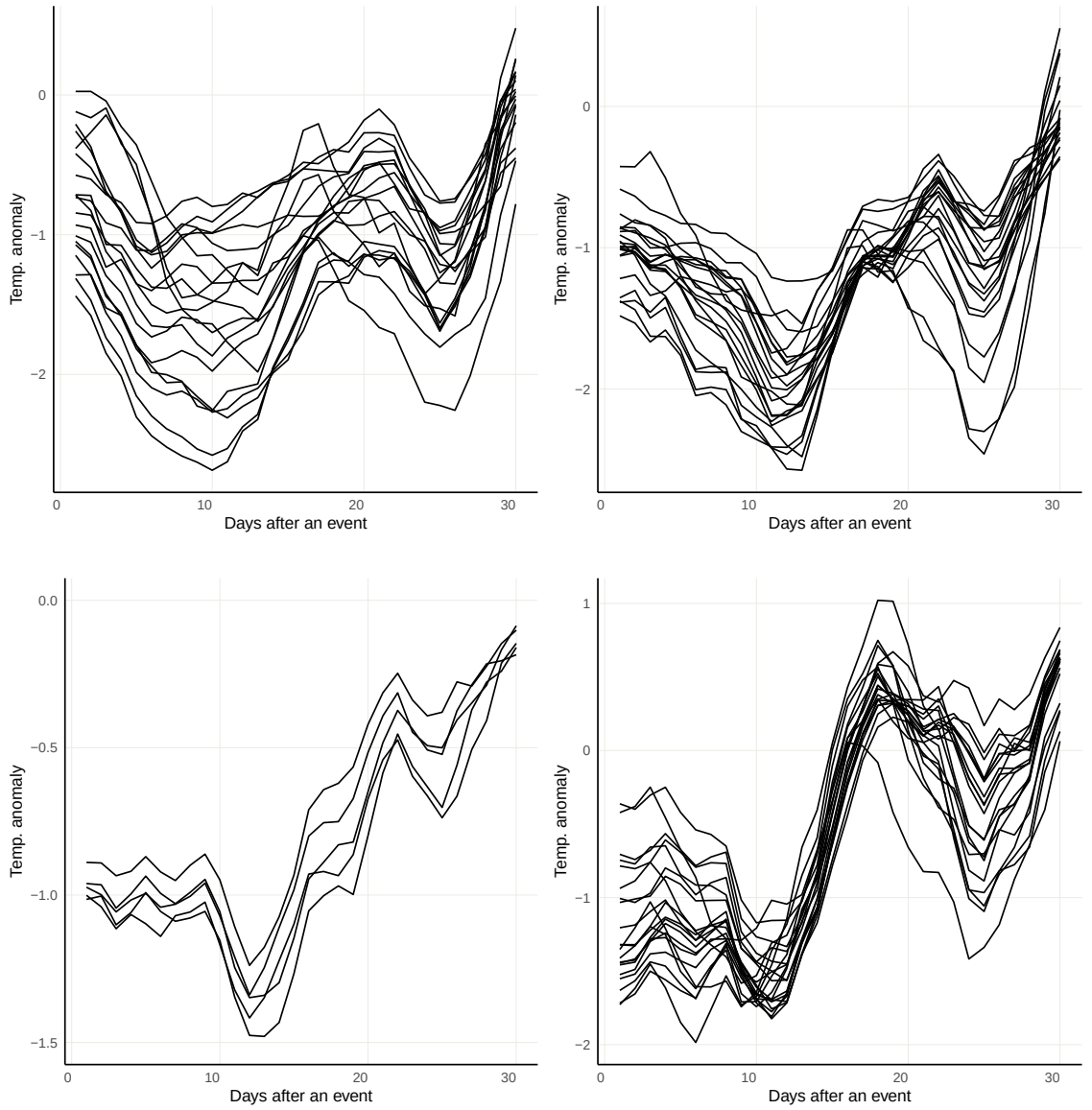

Figure 6. Impulse function of SSWs in (from top left to bottom right) Norway, Sweden, Denmark and Finland, showing the mean temperature anomaly with respect to climatology n days after an event was observed. Each line corresponds to a county in the respective country.

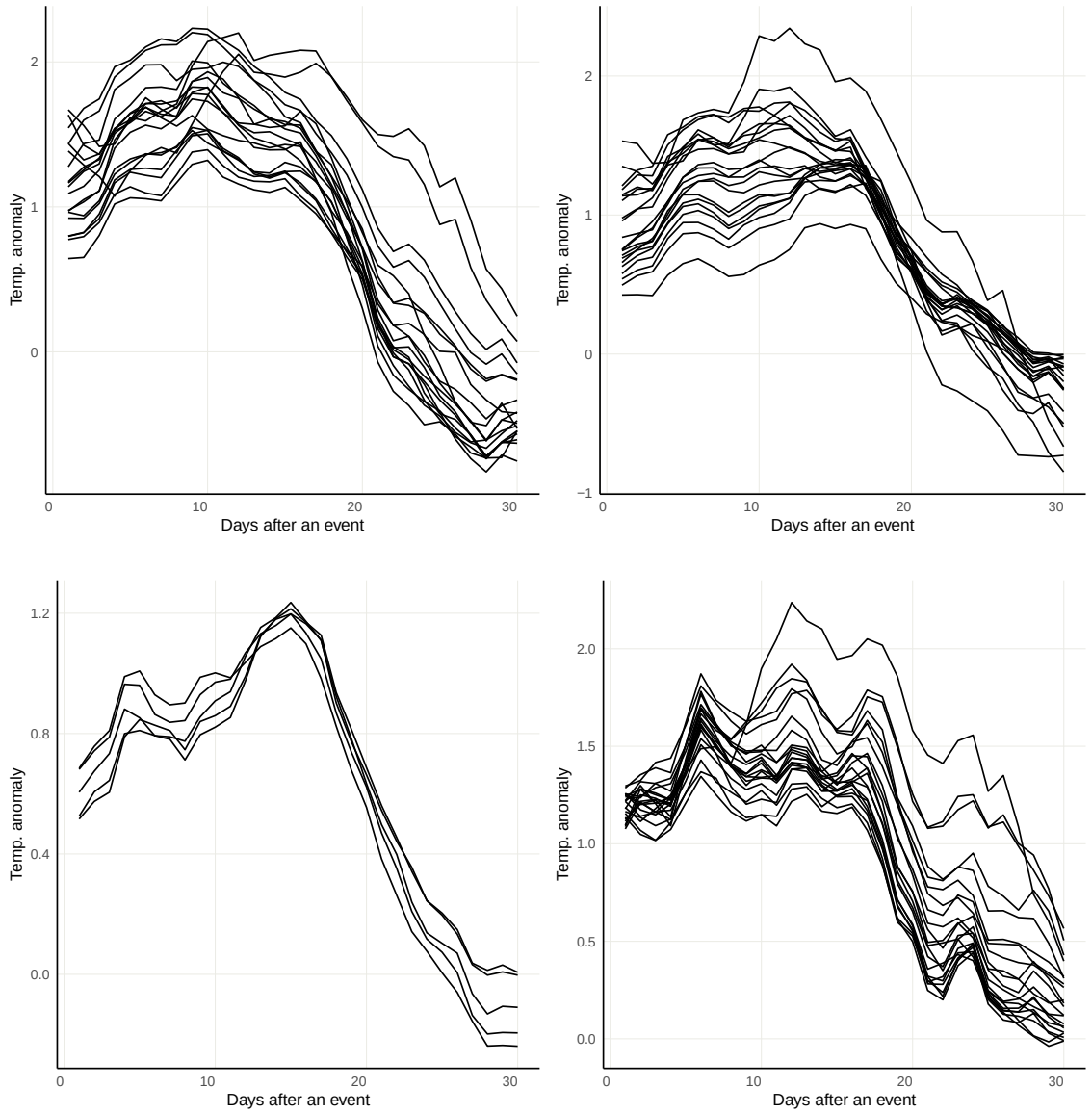

Figure 7. Impulse function of SPVs in (from top left to bottom right) Norway, Sweden, Denmark and Finland, showing the mean temperature anomaly with respect to climatology n days after an event was observed. Each line corresponds to a county in the respective country.

The mean daily MSE reduction, averaged over the first 14 and 21 days following a SPV event, in all Scandinavian counties is presented in figure 10 and 11. The MSE reduction is largest in the southern part of Norway and Sweden, as well as in Denmark, and the largest gains are in counties along the Oslo fjord. In Finland and the northern part of Sweden the impulse function of SPVs appears to have little to no skill in temperature prediction.

Following a SSW event, however, the story becomes more divisive. Figure 12 and 13 show the MSE in Oslo and Finnmark at lead times one to thirty days following an observation of a SSW event. The mean daily MSE reduction, averaged over the first 14 and 28 days following a SSW event, in all Scandinavian counties presented in figure 14 and 15, tells the same story: Counties bordering to Skagerrak and Kattegat enjoy a large MSE reduction, while Finland, North-Sweden and Norway's western and northern coast display no skill. And again counties arond the Oslo fjord enjoy the largest MSE reduction of about 10–12 % during the first four weeks following an observed SSW event.

4.7.1 Optimal event classification

While the SSW climatology appears to be inept in all Norwegian regions but Østlandet, recall, however, that a SSW event is somewhat randomly defined to occur when the zonal wind is less than $u = 0$ ms^{-1} – while said limit might have some precedence, it is too a limit choosen because of its roundness and simplicity. For a zonal wind y, such that an event occurs when $(y \leq u)$, can we find a better (optimal) limit $\tilde{u}$ in the regions where the SSW climatology performed poorly?

Regarding the definition of *optimiality*, there are two apparent choices: Find $\tilde{u}$ such that the average MSE gain is maximized over the first l days following all events $(y \leq \tilde{u})$, or find $\tilde{u}$ such that the the average MSE gain is maximized over all days in all winters. For simplicity we focus on the first approach and set $l = 21$ days.

See figure 16 for this approach applied to Finnmark, where the average $\tilde{u}$ was found to be $\bar{\tilde{u}} = 17.09$ ms^{-1}. Over the first three weeks following a SSW-like event, event climatology now performs 0.83 % better than climatology on average, while it before was 5.11 % worse. This event climatology still retains a modest skill compared to that of Østlandet, but this custom SSW-like event is more frequently occuring than SSWs, and it is consequently a skill that might be used during more winters.

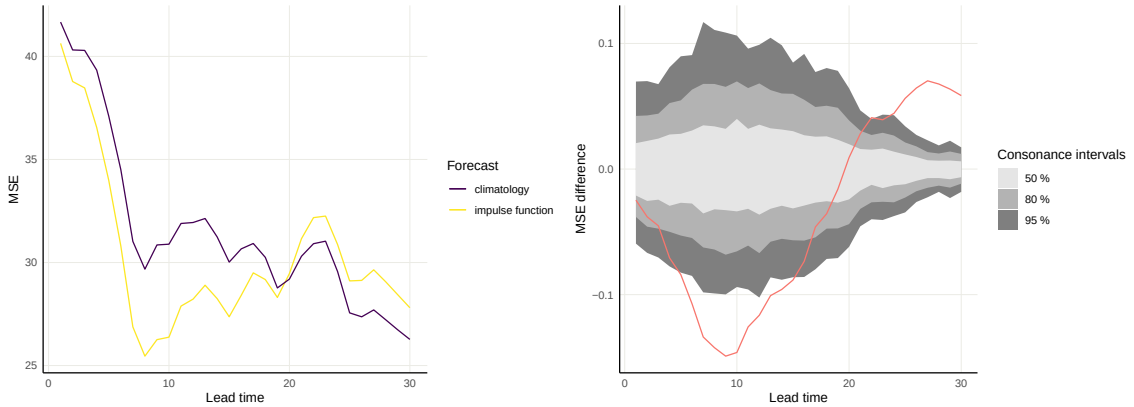

Figure 8. Mean squared prediction error in Oslo for climatology (purple) and SPV climatology based on the impulse function of SPVs (yellow), for lead times of one to thirty days ahead after a SPV event was observed. The right panel shows the result of a permutation test on the significance of the difference between event climatology and temperature climatology at each lead time.

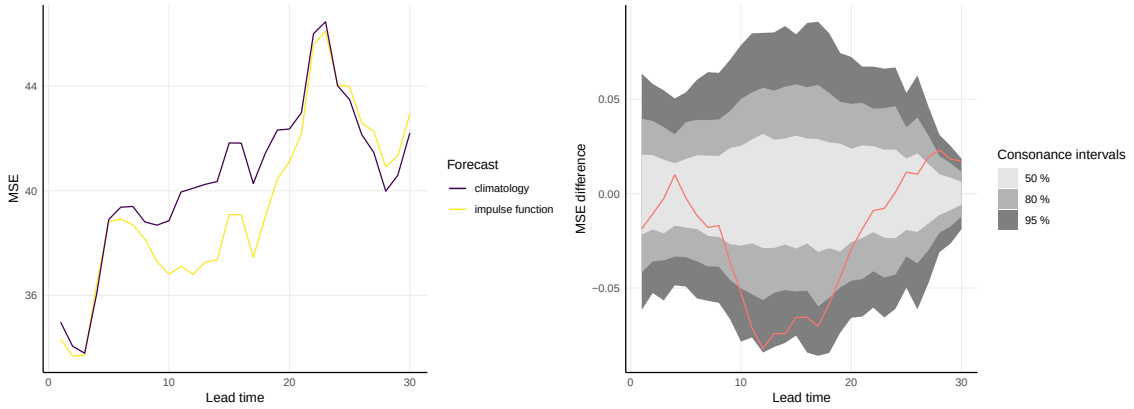

Figure 9. Mean squared prediction error in Finnmark for climatology (purple) and SPV climatology based on the impulse function of SPVs (yellow), for lead times of one to thirty days ahead after a SPV event was observed. The right panel shows the result of a permutation test on the significance of the difference between event climatology and temperature climatology at each lead time.

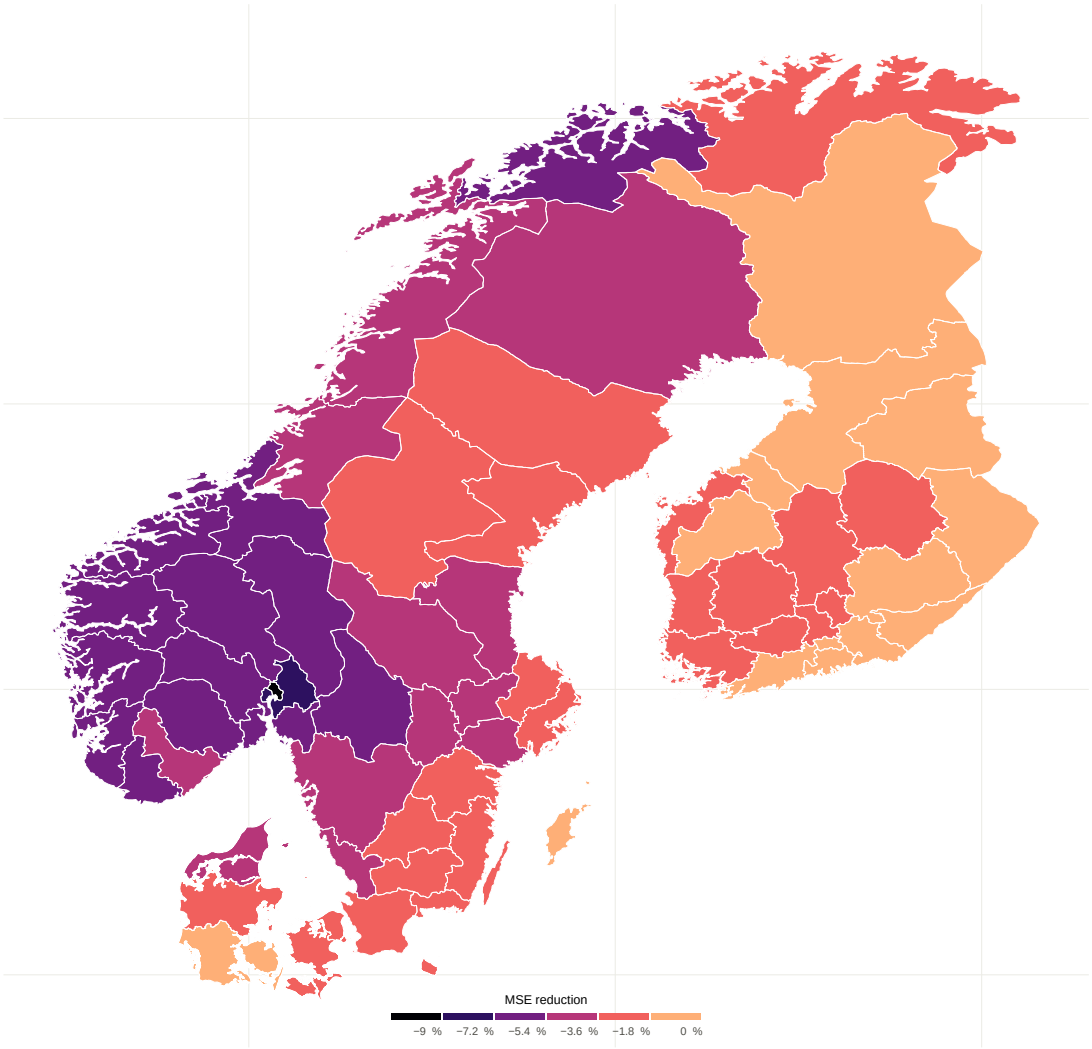

Figure 10. Reduction in daily mean squared prediction error, averaged over the first 14 days after a SPV event was observed, relative to climatology in each Norwegian/Swedish/Finnish county, when 1 to 14 days ahead was forecasted based on the impulse function of SPVs following the observation of a SPV event.

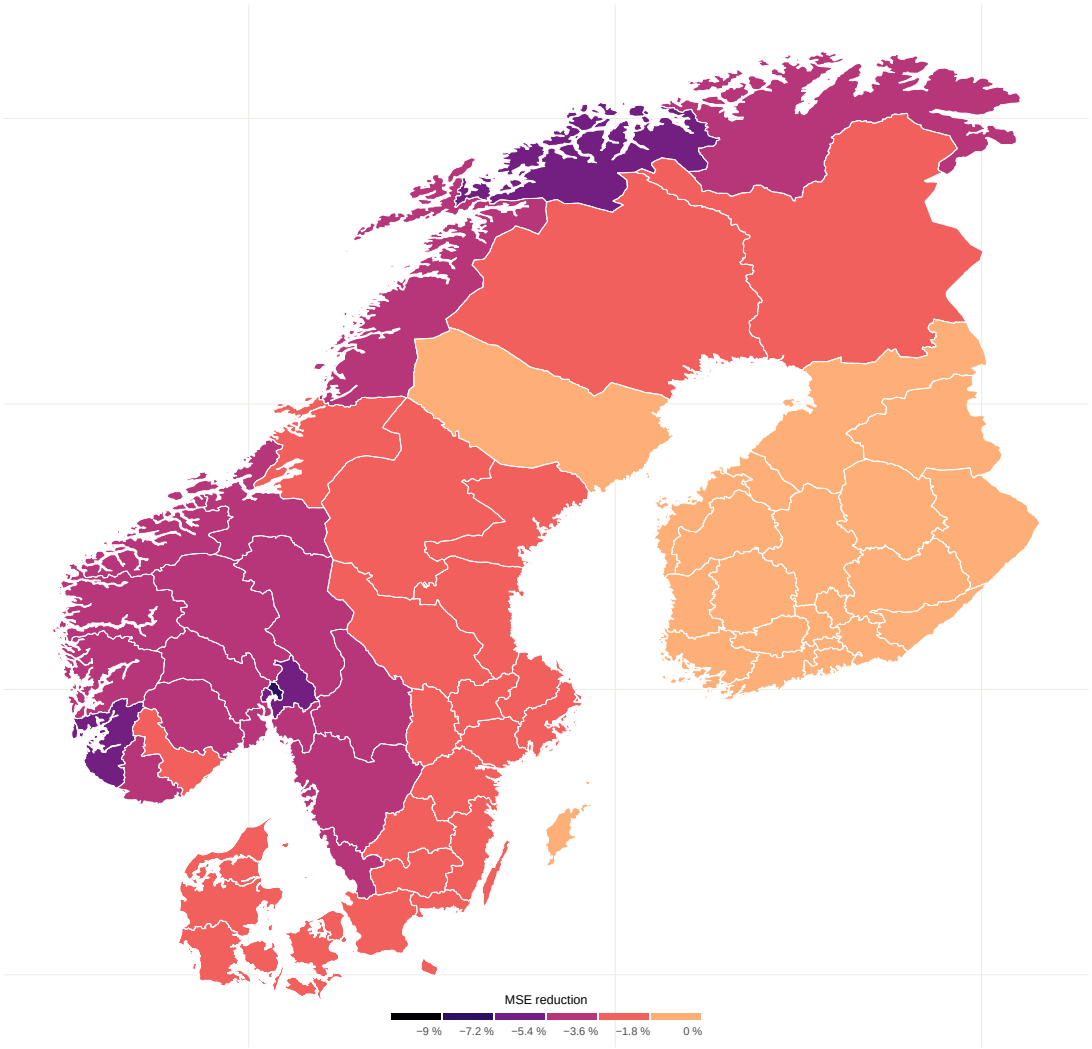

Figure 11. Reduction in daily mean squared prediction error, averaged over the first 21 days after a SPV event was observed, relative to climatology in each Norwegian/Swedish/Finnish county county, when 1 to 21 days ahead was forecasted based on the impulse function of SPVs following the observation of a SPV event.

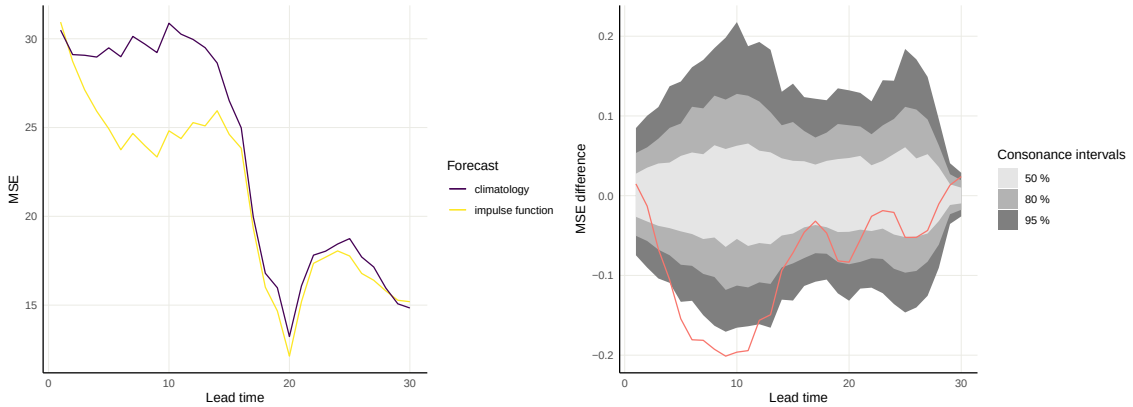

Figure 12. Mean squared prediction error in Oslo for climatology (purple) and SSW climatology based on the impulse function of SSWs (yellow), for lead times of one to thirty days ahead after a SSW event was observed. The right panel shows the result of a permutation test on the significance of the difference between event climatology and temperature climatology at each lead time.

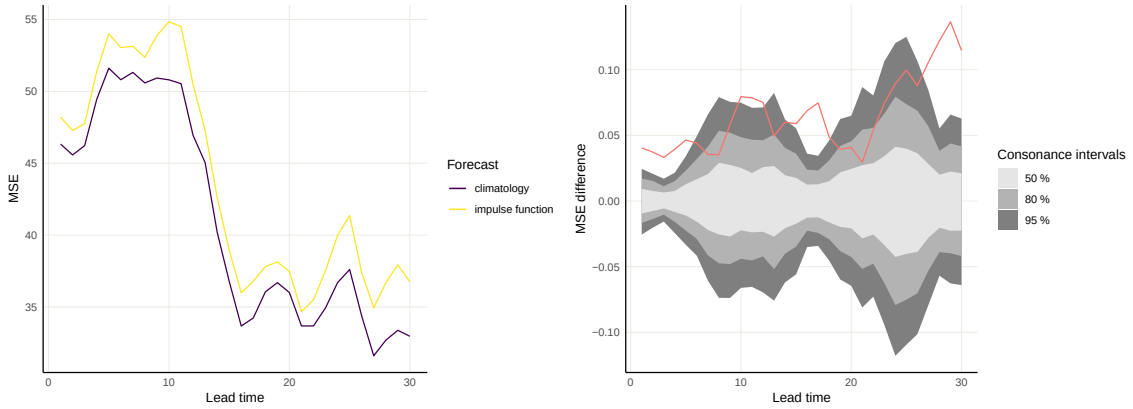

Figure 13. Mean squared prediction error in Finnmark for climatology (purple) and SSW climatology based on the impulse function of SSWs (yellow), for lead times of one to thirty days ahead after a SSW event was observed. The right panel shows the result of a permutation test on the significance of the difference between event climatology and temperature climatology at each lead time.

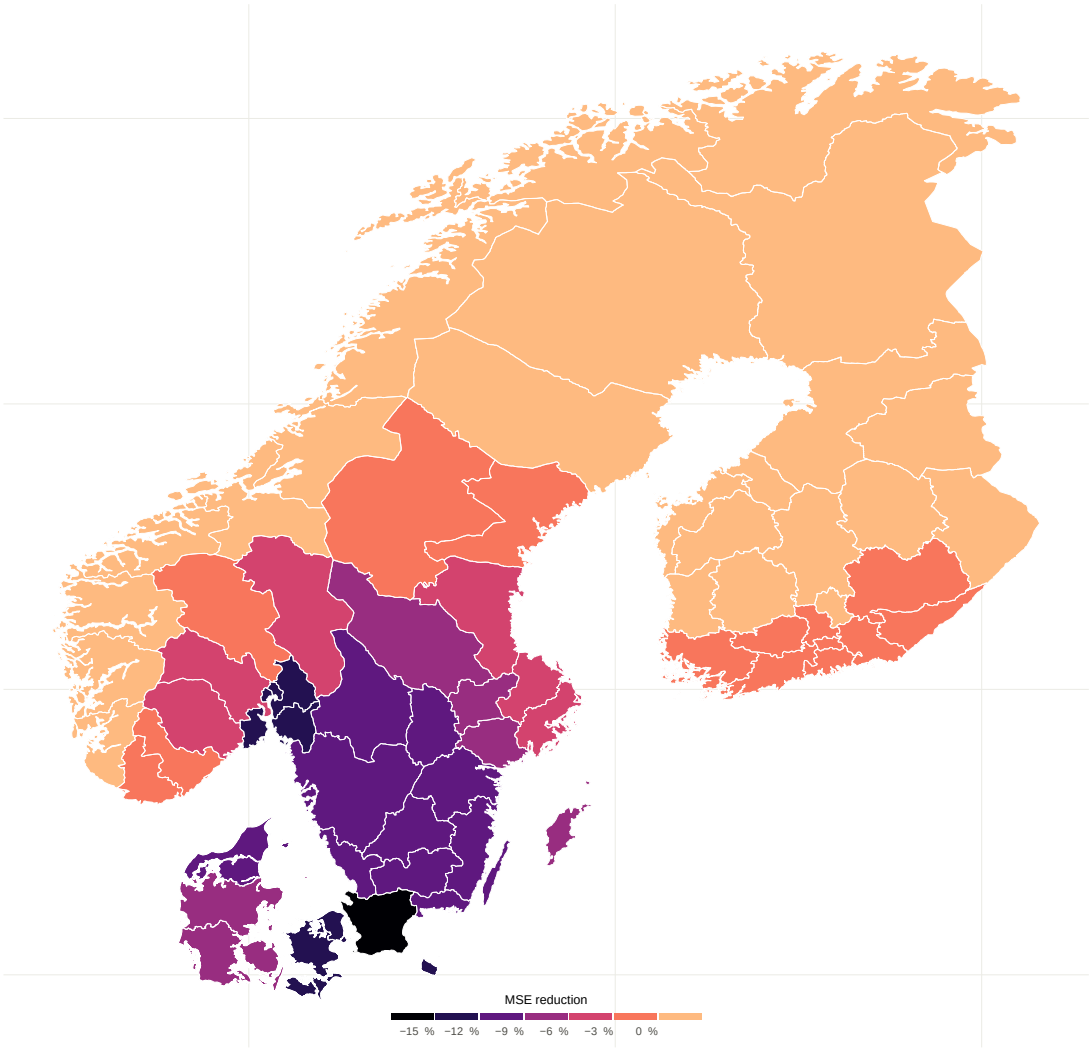

Figure 14. Reduction in daily mean squared prediction error, averaged over the first 14 days after a SSW event was observed, relative to climatology in each Norwegian/Swedish/Finnish county county, when 1 to 14 days ahead was forecasted based on the impulse function of SSWs following the observation of a SSW event.

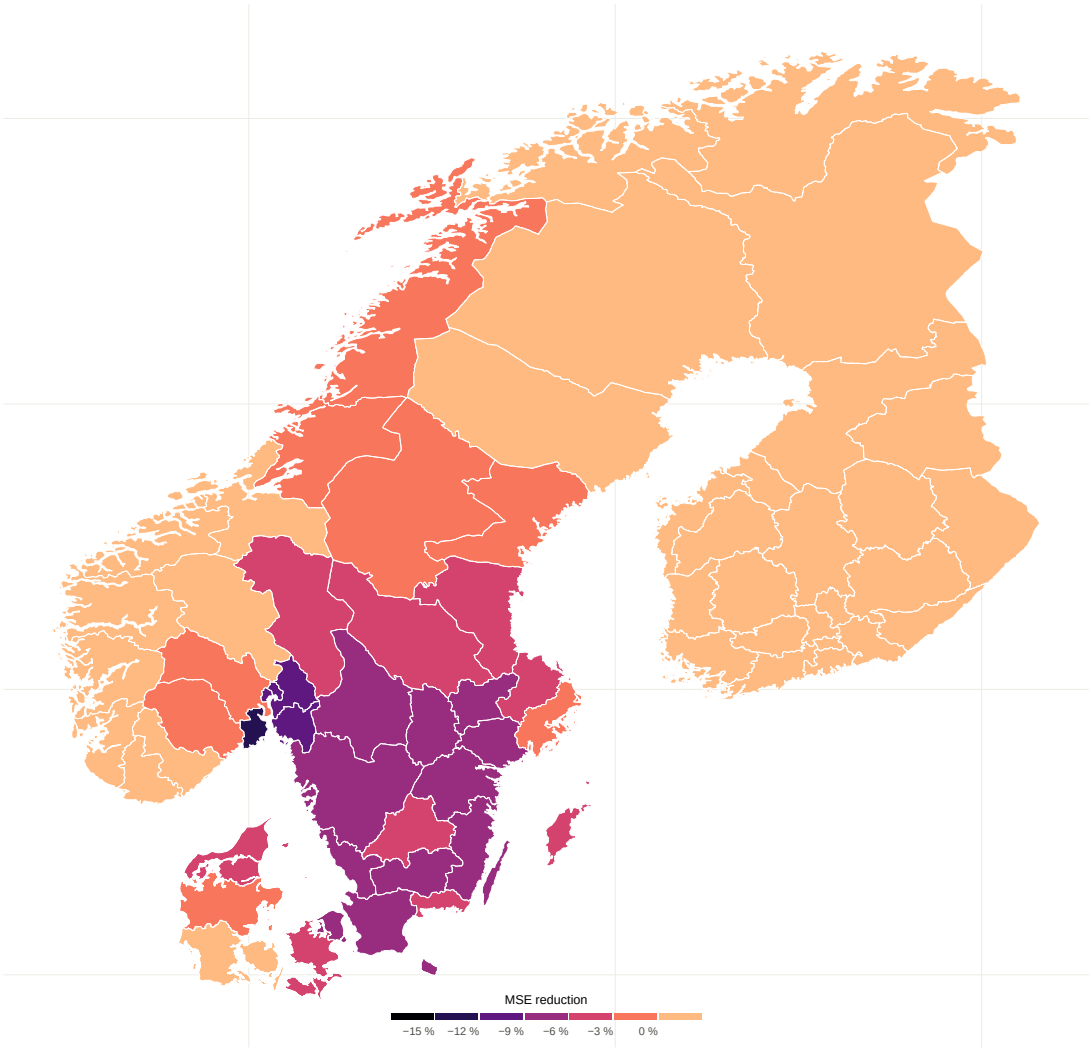

Figure 15. Reduction in daily mean squared prediction error, averaged over the first 28 days after a SSW event was observed, relative to climatology in each Norwegian/Swedish/Finnish county county, when 1 to 28 days ahead was forecasted based on the impulse function of SSWs following the observation of a SSW event.

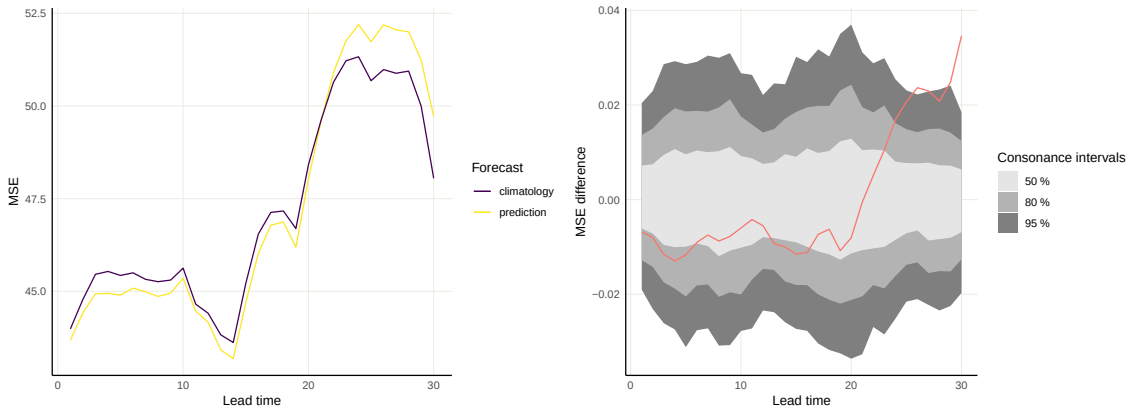

Figure 16. Mean squared prediction error in Finnmark county for climatology (purple) and SSW climatology based on the impulse function of SSWs (yellow), for lead times of one to thirty days ahead after a SSW-like event was observed. The right panel shows the result of a permutation test on the significance of the difference between event climatology and temperature climatology at each lead time.

4.8 Regression of temperature anomaly on u wind

Instead of merely focusing on the (somewhat arbitrarily defined) occurance of discrete SSW and SPV events, we wish to assess the correlation between the zonal mean U wind and winter surface temperatures in Norway.

We first fit a simple linear regression in each Norwegian county, where we regress the temperature anomaly on day $d_0 + l$, for lead time l, with respect to climatology on the zonal mean U wind anomaly over the previous d days with respect to climatology. That is, let the mean temperature in region r_0 at year y_0 and day of winter d_0 be t_{r_0,y_0,d_0}, and let the mean uwind for the previous d days be u_{d,y_0,d_0}. Let climatology for temperature be $\overline{t_{r_0,y_0,d_0}}$ and define climatology for mean uwind for the previous d days by $\overline{u_{d,y_0,d_0}} = \frac{1}{Y-1}\sum_{y\neq y_0} u_{d,y,d_0}$. We thus want to fit the model

$$t_{r_0,y_0,d_0+l} - \overline{t_{r_0,y_0,d_0+l}} = \beta_1(u_{d,y_0,d_0} - \overline{u_{d,y_0,d_0}}) + \varepsilon, \tag{16}$$

where $\varepsilon \sim \mathcal{N}(0,\sigma^2)$, for all regions r_0.

We have, however, observed that (predominantly) large U wind anomalies have an effect on surface temperatures. Consequently we too want to fit a GLM with a Gaussian family and a link function which flattens out the effect of small wind anomalies. That is, assume $Y_i \sim N$ with $\mathrm{E}Y_i = \mu_i$ and $\mathrm{Var}Y_i = \sigma^2$, for $i = 1,\ldots,n$. Let the systematic component be $\eta_i = \mathbf{x}_i^T\boldsymbol{\beta}$ with link function $\mu_i = g^{-1}(\eta_i) = (\eta_i/\alpha)^3$, for $\alpha > 1$. This gives a log-likelihood function

$$l(\boldsymbol{\beta}) = -\frac{n}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2, \tag{17}$$

and a score function

$$\mathbf{s}(\boldsymbol{\beta}) := \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{3}{\sigma^2 \alpha^3} \sum_{i=1}^{n} (y_i - \mu_i) \eta_i^2 \mathbf{x}_i. \tag{18}$$

The expected Fisher information matrix is thus

$$\mathbf{F}(\boldsymbol{\beta}) := \mathrm{Cov}(\mathbf{s}(\boldsymbol{\beta})) = \frac{9}{\sigma^2 \alpha^6} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^T \eta_i^4, \tag{19}$$

and $\widehat{\boldsymbol{\beta}}$ is found using a Fisher-scoring method, cf. $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \mathbf{F}(\boldsymbol{\beta}^{(t)})^{-1} \mathbf{s}(\boldsymbol{\beta}^{(t)})$. Equivalently we too wish to fit a polynomial regression

$$t_{r_0,y_0,d_0+l} - \overline{t_{r_0,y_0,d_0+l}} = \beta_1 (u_{d,y_0,d_0} - \overline{u_{d,y_0,d_0}})^3 + \varepsilon. \tag{20}$$

Fitting the three aforementioned models and estimating their error by LOOCV, their MSE reductions with respect to the MSE of climatology, for different configurations of (d, l) in Oslo and Troms, is presented in figure 17. Oslo temperatures appear to be mostly influenced by the larger, extreme wind anomalies, while temperatures in northernmost counties such as Troms and Finnmark is best described by the simple linear regression.

The maximum daily MSE reductions in Oslo and Troms is at a modest 1.5 % for optimal choices of (d, l). We could still, in theory, combine this analysis with the UKMO or ECMWF zonal mean U wind forecasts in order to improve temperature forecasts in the two–three first weeks of December where the U wind forecasts had a significant skill over climatology.

Due to only having 1st November-initialised forecasts from 1st December available, we attempt to predict Oslo temperatures during the third and fourth week of December using the polynomial regression in (20) with $(d = 10, l = 3)$. This gives a 0.37 % daily MSE reduction compared to climatology. The lack of improvement is unsurprising: The maximum MSE reduction of 1.5 % over climatology in figure 17 was achieved when using true zonal mean U wind observations; using somewhat askew forecasts will necessarily reduce said predictive power.

4.9 Improving long-range temperature forecasts by impulse functions and Cox regression

Section 4.7 concerning state-dependent event climatology forecasts demonstrated that climatological surface temperature forecasts could be vastly improved, in certain regions, in the wake of SSW and SPV events. Consequently, using UKMO and ECMWF zonal mean U wind forecasts, could impulse functions be used to improve temperature forecasts on the time scale of one–two months ahead in time?

Using each ensemble member to compute the temperature anomaly at a given lead time, and then taking the mean over all ensemble members, the MSE difference between the

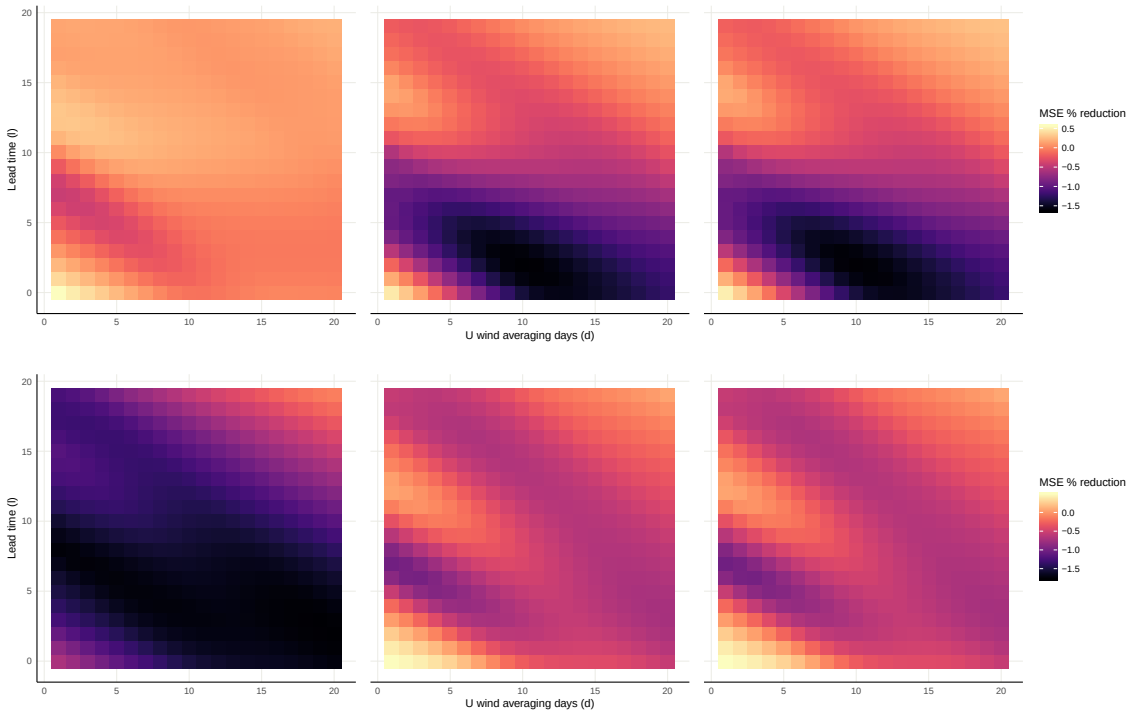

Figure 17. Percentage-wise MSE reduction with respect to climatology, for different configurations of (d, l), for (from left) a simple linear regression, a GLM with a Gaussian family and a link function which flattens out the effect of small wind anomalies, and a polynomial regression of degree three. Top row: Oslo, bottom row: Troms.

impulse function-based forecast and climatology at each lead time is presented in figure 18, for zonal mean U wind forecasted by UKMO and ECMWF. Since neither forecaster manages to accurately predict when an event will occur on the time scale of one–two months ahead in time, the impulse function-based temperature forecasts perform slightly worse than climatology.

UKMO and ECMWFs inability to accurately predict when an event will occur could mayhap be remedied by fitting a Cox regression to the data, such that we at each lead time could predict the hazard of an event occuring based on the forecasted zonal mean U wind. One fitted model, a Cox regression with $(u_{y_0,d_0} - \overline{u_{y_0,d_0}})^3$ (the cube of mean zonal U wind minus climatology) as a covariate, proved significant at the 1 % level. (We could also attempt to fit a logistic regression to the data.) It is remains, however, to combine the Cox regression with the impulse function-based forecasts.

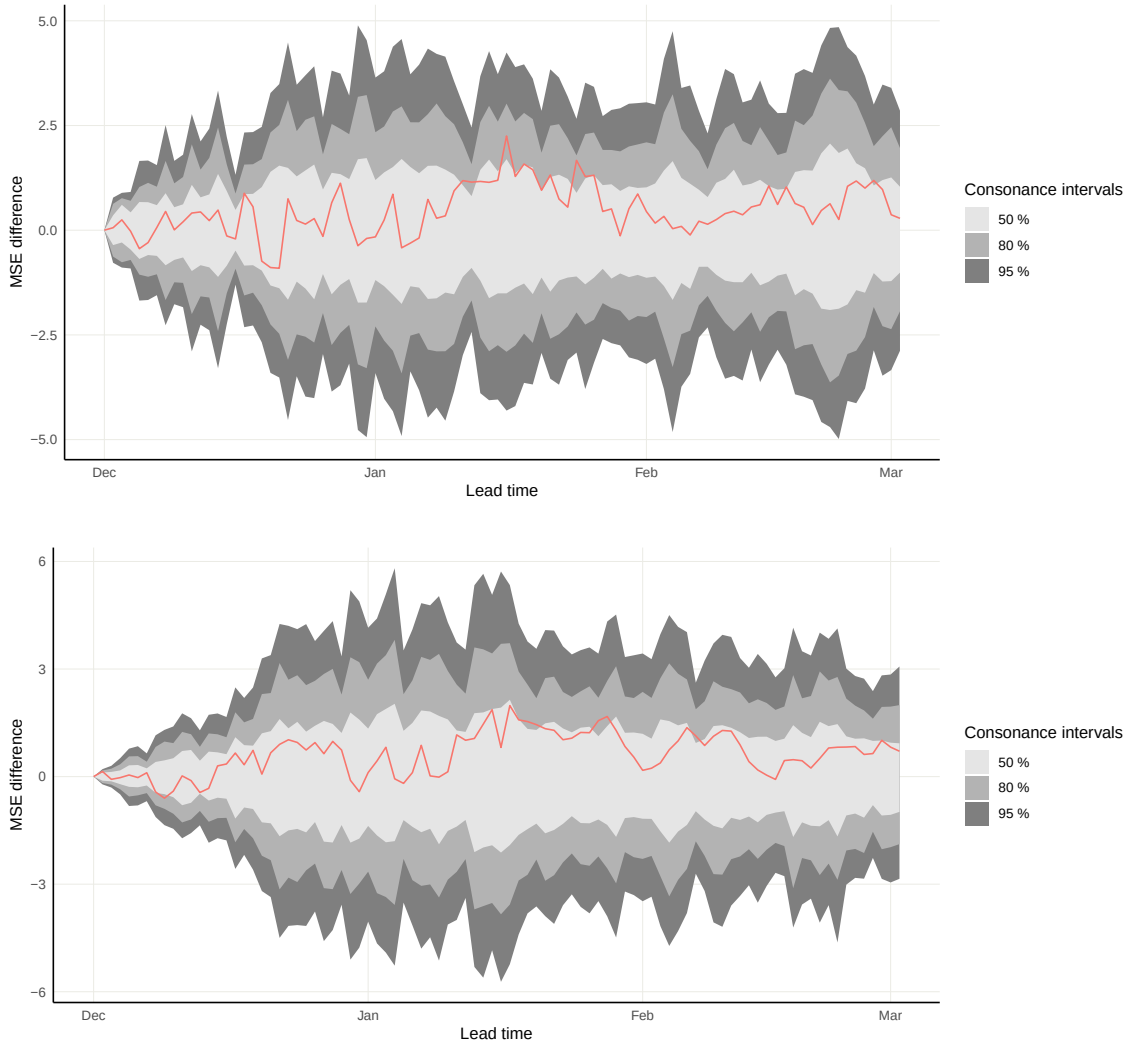

Figure 18. Permutation tests at each lead time of the difference between the MSE of impulse function-based temperature forecasts and the MSE of climatology, where the impulse functions is based on the forecasted zonal mean U wind from UKMO (top) and ECMWF (bottom).

5 Conclusion

When conditioning Scaife's perfect probabilities on actual events instead of proxy observations, the UKMO ensemble forecast of the stratospheric polar vortex lost all its apparent skill in predicting strong polar vortices on the time scale of months. While UKMO had some skill in differentiating the probability of sudden stratospheric events occuring between years in which such events did and did not occur, the Brier score of UKMO was not found to be significantly different from that of climatology on the time scale of months. ECMWF was not found to be significantly different from climatology in predicting the occurance of stratospheric events.

Stratospheric events appear to cause significant La Niña and El Niño-like weather patterns on the time scale of two–four weeks following an event, where the most pronounced effects were in regions bordering to the Oslo fjord, Skagerrak, Kattegat and the lower Baltic sea. Creating a non-paramatric temperature forecast based on this relationship proved to greatly improve temperature forecasts in some regions, but due to both UKMO and ECMWF displaying limited skill in predicting the occurance or the timing of occurance of stratospheric events, it proved difficult to combine the non-parametric weather patterns following a stratospheric event with forecasts of the stratospheric circulation.

Regressing temperature anomaly at some lead time on the zonal mean U wind averaged over a previous period of time proved to explain some of the variation in winter surface temperature, but this analysis too struggled to significantly improve long-range temperature forecasts when using the long-range zonal mean U wind from UKMO and ECMWF.

References

Baldwin, M. P. and Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, 294(5542):581–584. 5

Friederichs, P. and Thorarinsdottir, T. L. (2012). Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, 23(7):579–594. 8

Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media. 8

Möller, A., Lenkoski, A., and Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, 139(673):982–991. 8

Murphy, A. H. (1993). What is a good forecast? an essay on the nature of goodness in weather forecasting. *Weather and forecasting*, 8(2):281–293. 7

Scaife, A., Karpechko, A. Y., Baldwin, M., Brookshaw, A., Butler, A., Eade, R., Gordon, M., MacLachlan, C., Martin, N., Dunstone, N., et al. (2016). Seasonal winter forecasts and the stratosphere. *Atmospheric Science Letters*, 17(1):51–56. 4, 5, 10, 11