Response to comments of Reviewer 3

We thank the reviewer for providing detailed comments (shown inside double quotes below).

Major comments:

1) "Page 3522, Title of the manuscript: '... winter extreme precipitation...' can be modified as '... winter 5-day extreme precipitation...' "

   After weighing the trade-off between a longer and a shorter title, we decided to keep the original shorter title, as it was already quite long, and pentad (5-day) total precipitation is commonly used.

2a) "Page 3522, abstract, line 8: '... the winter extreme precipitation were the predictands.' What is predictand in the study? Is it 'precipitation' or 'PCs extracted from 5-d precipitation anomaly'?"

   The predictands are the leading PCs of the winter season's maximum 5-day total precipitation anomalies.

2b) "Page 3522, Abstract, lines 13-15: '... two robust SVR models tended to have better forecast skills than the two non-robust models (MLR and BNN),...'
   Is the result from the study unexpected, or is it obvious? Is the word robust used in the right context? Can authors mention differences between robust, effective and efficient models? If it is already known that some models are robust and some models are non-robust, what is the need to make comparison between them?"

   The mean absolute error (MAE) norm is considered to be robust to outliers in the data. As the error norm used in SVR is rather similar to the MAE norm, SVR is also considered to be robust. In some of our other studies, e.g. Arctic blizzard wind forecasts, SVR did not improve on BNN, so it is not obvious that a robust method will out-perform a non-robust method. The mean squared error (MSE) norm used by BNN is in fact optimal for Gaussian noise. Extreme data, such as the winter season's maximum 5-day total precipitation anomalies, are non-Gaussian.

2c) "Page 3522, abstract, line 16: '... Among the six regions, the Eastern...'
   Which six regions are being referred?"

   The six regions refer to the six geographic regions over Canada as mentioned in line 7 above.

3) "Where are the values of parameters estimated for each of the models considered in the study? They have to be given in the manuscript for obvious reasons."

   It is impractical to list the parameters used in our models. For instance, each single BNN model has of the order of 20 parameters. There are 30 ensemble members for

4 forecast lead times over 6 geographic regions for about 5 PC predictands in each region. Hence for BNN, there are of the order of $20 \times 30 \times 4 \times 6 \times 5 = 7 \times 10^4$ parameters. Furthermore, the cross-validation process means models were trained and their forecast ability measured over a 3-year validation period. Over the whole record, there are 19 such validation periods, so one has to multiply 7×10^4 by 19. Also parameters used in NN models are not easily interpretable.

4) "Page 3526, lines 6-10: 'The reason that the maximum 5-d total precipitation instead of the daily extreme is used here because ... heavy precipitation are mostly due to multi-day episodes. Maximum 5-d precipitation has been also chosen as one of the standard seasonal extreme precipitation indices by the European Union STARDEX project'
(a) The choice of 5-day precipitation for the study lacks proper reasoning. It is agreed that multi- day episodes might cause floods, but multi-day need not necessarily indicate 5-day period. Did the authors prepare any frequency plots for the data being analyzed to identify low-frequency signal? If so, the plot must be presented in the manuscript to justify considering analysis of only 5-day episode as adequate for the study area.
(b) What are the other standard seasonal extreme precipitation indices chosen by STARDEX project? Why they are not considered in the present study?"

The reason we used the 5-day total precipitation is that it follows a standard index used for extremes in the large European project STARDEX. We do not expect the results to change significantly if we use say 4-day total precipitation, but then it would not be a standard index, making our results difficult to compare with those by other researchers. Other extreme indices are given on p.5 in the STARDEX final report, downloadable from the site `http://www.cru.uea.ac.uk/projects/stardex/`. Strictly speaking, the STARDEX index was for rainfall instead of precipitation. However, 5-day (pentad) precipitation data is widely used, e.g. SSM/I pathfinder pentad data (`http://disc.sci.gsfc.nasa.gov/precipitation/documentation/readme_html/ssmi_monthly_readme.shtml`), and
Global Precipitation Climatology Project pentad precipitation data (`http://islscp2.sesda.com/ISLSCP2_1/html_pages/groups/hyd/gpcp_precip_pentad_xdeg.html`).

5) "Page 3522, last line: '... the long-term trend of extreme precipitation events seems not so significant in most areas of Canada (Zhang et al., 2001; Kunkel, 2003)...' The statement 'seems not so significant' is subjective. Though there is growth in evidence of climate change on extreme precipitation in different parts of the world, it is surprising to note that there is no trend in the Canadian precipitation data, over the period (1950 to 2006) being considered for the analysis. What are the tests and significance level considered by authors to test the long-term trend?"

Zhang et al. (2001) mentioned that the observed increase in precipitation totals was mainly due to increase in the number of small to moderate events, and Kunkel (2003) expressed the similar opinion. In Zhang's study, they calculated the linear trends,

their statistical significance, and their 95% confidence intervals for various indices (heavy precipitation frequency and intensity, and percentiles of daily precipitation) using a Kendall's tau-based nonparametric procedure. The statistical significance of correlations was evaluated by the F test. Significance in trends and in correlations were assessed at the 5% level. Zhang et al. concluded in their Section 5: 'The lack of generally increasing trend in heavy and extreme events over the last century in this high-latitude country fails to support GCM projections, possibly due to the relatively early stage of greenhouse gas–induced global warming.'

6) "Page 3523, lines 13-15: '... Most seasonal forecasts focus on predicting the seasonal mean of the precipitation instead of seasonal statistics of extreme precipitation events.'
(a) What do authors mean by 'seasonal mean of the precipitation'? Should it be 'seasonal cumulative precipitation'?
(b) Which seasonal statistics of extreme precipitation are being referred? Are they statistics such as mean, standard deviation, skew, kurtosis, or is the reference to accumulated precipitation over several days (3-day, 5-day, 7-day etc.). It would be better if authors can list potential uses of predicting the seasonal statistics of extreme precipitation events."

(a) The two are the same up to a constant. Seasonal mean is the seasonal cumulative amount divided by the number of days in a season.
(b) See the STARDEX extreme indices mentioned in (5). Seasonal extreme forecasts will be useful e.g. to the agricultural sector.

7) "Page 3524, lines 19-20: '... the predictand is the very noisy and non-Gaussian winter extreme precipitation anomaly.' There is inconsistency in definition of predictand (see abstract, line 8: '... the winter extreme precipitation were the predictands'). Is the predictand maximum winter 5-day accumulated precipitation, or precipitation anomaly, or principal components (PCs) extracted from precipitation anomaly?"

In the Abstract and in the Introduction, we did not want to burden the reader with too much detail, so our description of the predictand was simplified. Only later did we specify that the predictands were the leading PCs extracted from the precipitation anomaly. We will avoid ambiguity in the revised manuscript.

8) "Page 3525, lines 7-8: '... removing the climatological cycle from the monthly mean data and filtering them using a 3-month running mean...' The write-up lacks clarity. It would be better if equation is given to explain this part of the analysis. Does 'removing the climatological cycle' indicate deseasonalization (or removing periodicity) in data? Or does it indicate removing long-term cyclicity in the data? What is the reason for choosing 3-month running mean for filtering?"

Removing the climatological cycle means deseasonalization. We will clarify this in the revised manuscript. The 3-month running mean generates 3-month (seasonal) mean values from monthly values.

9) "Page 3525, line 9: '. . . After normalizing the anomalies, time-lagged copies of the data were stacked. . . '. How is the data normalized? and why is it normalized? If deseasonalization is done in the first step, then perhaps normalization is not necessary."

The anomalies at each gird were normalized (a better term is 'standardized') by dividing by their standard deviation. We agree the normalization is not necessary (but does no harm). [Earlier in the research, we assembled the SST and Z500 anomalies together before doing the space-time PCA, and in that case, the normalization step was necessary].

10) "Page 3526, lines 3-4: 'The climatological seasonal cycle of 5-d precipitation was then removed, and the 3-month maximum was identified as the seasonal extreme precipitation anomaly.:' The statement lacks clarity, and it would be better if equation is given to explain this part of the analysis. How was the climatological seasonal cycle removed? Does it indicate deseasonalization?"

The climatological seasonal cycle is the average of the maximum 5-day total precipitation for each calendar month over the years 1950-2007. It was removed from the monthly maximum 5-day total precipitation to get the monthly extreme precipitation anomaly. Then the 3-month maximum of the anomaly is identified as the seasonal extreme precipitation anomaly.

11) "Page 3525, line 14: '. . . This PCA process, ..., is performed on the SST and Z500 normalized anomalies separately, each having 5 leading principal components (PC) retained. . . '
Principal components must be PCs, rather than PC. What is the logic behind choosing 5 PCs? How much variance did they preserve?"

In unpublished research by Dr. Aiming Wu, he varied the number of EEOF PCs in neural network predictions of North American seasonal climate and found that using more than 5 leading PCs did not further improve forecast skills. The variance explained by the 5 PCs of SST was 52%, while the variance explained by 5 PCs of Z500 was 37%.

12) " '. . . In view of the diversity of the Canadian climate, we classified the 118 stations into six groups using K-means clustering. . . ' (a) The description of K-means cluster analysis lacks clarity. (b) How is it decided that there are six groups? Cluster validity measures have to be used. Authors can refer some of the latest works in hydrology (e.g., Rao and Srinivas, 2006) to know the procedure. (c) How many feature vectors form input to K-means clustering algorithm, and what are the elements in each feature vector?"

We tried different numbers of clusters from 2–8, and 6 was subjectively chosen because of its spatial consistency and clear physical/geographical interpretation. We will cite the more objective method by Rao and Srinivas (2006) in the revised paper. For each of the 118 stations, we compute the correlation of its seasonal extreme precipitation

anomaly with that of all 118 stations and the 6 climate indices, i.e. we use a total of 118+6 = 124 elements in each feature vector and there are 118 feature vectors in the cluster analysis.

13) "The data used for the study should be made available from a public domain (at least to reviewers), to allow for verifying reproduction of the results."

We will be happy to send data to reviewers via the journal editor if requested.

14) "Figures (a) Figure 1: What is the scale of the figure?
(b) Figures 2 to 7: The figures show 'average skill score' over all stations for each region. The 'average skill score' is not interesting. Instead, box-plots could have been presented for each region to draw correct inference about the range of skill scores (maximum, 95%, 75%, 50%(median), mean, 25%, 5% and minimum skill score) computed for each region. For each lead time, one box plot can be prepared for each model using estimates of skill scores for all the sites in the region."

(a) Scale added to the Fig.1.
(b) Figs.2–7 have been redrawn as boxplots.
Figs.8–10 have also been redrawn with slightly better colour scheme.

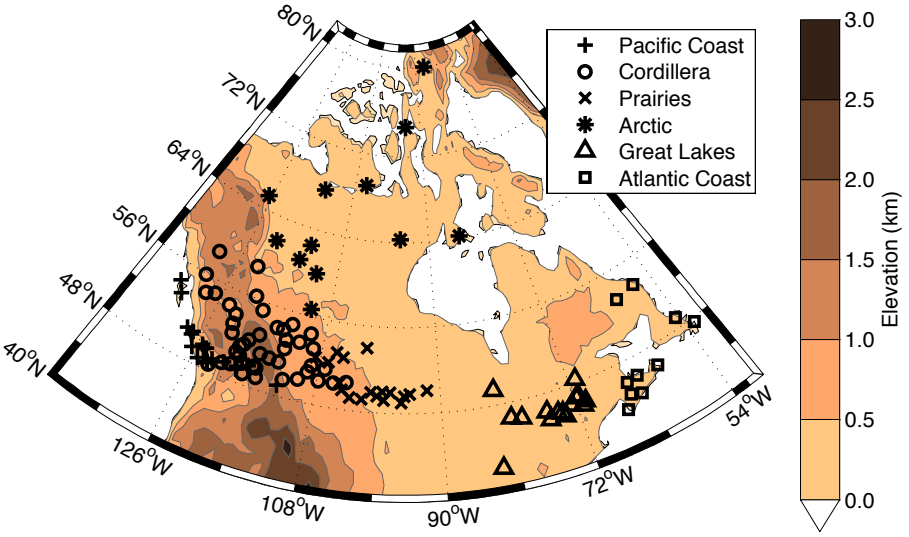

Figure 1: Spatial distribution of the Canadian stations, with different symbols used to indicate the six geographic regions determined by a cluster analysis. The shading illustrates the Canadian topography.

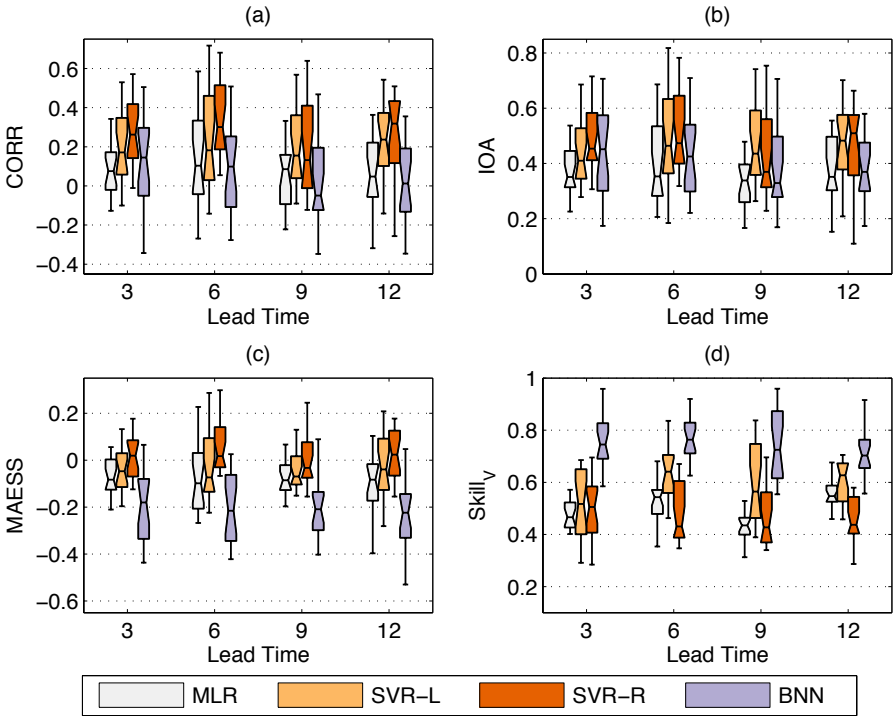

Figure 2: Cross-validated forecast scores, **(a)** CORR, **(b)** IOA, **(c)** MAESS and **(d)** Skill_V, in the Pacific coastal region (region R1) for the winter extreme precipitation at lead times of 3, 6, 9 and 12 months using the MLR, SVR with linear kernel (SVR-L), nonlinear SVR with RBF kernel (SVR-R) and BNN models. Lead time of 3 months means that predictor data up till September–November were used to forecast the December–February extreme precipitation. The "waistline" of the boxplot shows the median of all the values in the region R1, while the 25th and 75th percentiles are shown, respectively, as the bottom and top of each box. Distance between the 25th and 75th percentiles is the interquartile range (IQR). Data points more than 1.5 IQR from the median are considered outliers, and are shown as small circles (none in this figure). The whiskers indicate the data point nearest but not exceeding 1.5 IQR from the median.

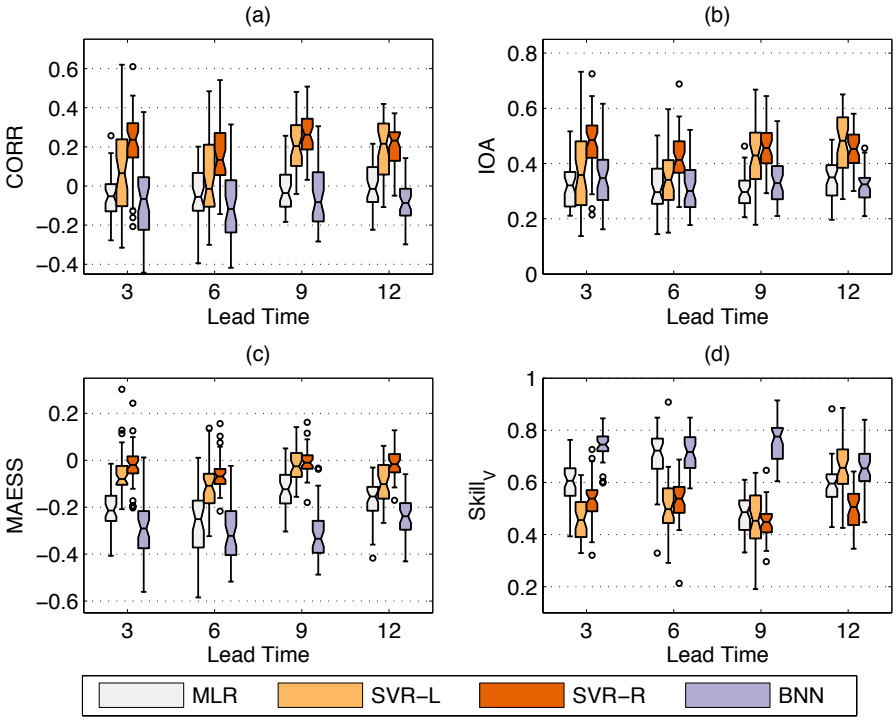

Figure 3: Same as Fig. 2, except over the Cordillera (region R2).

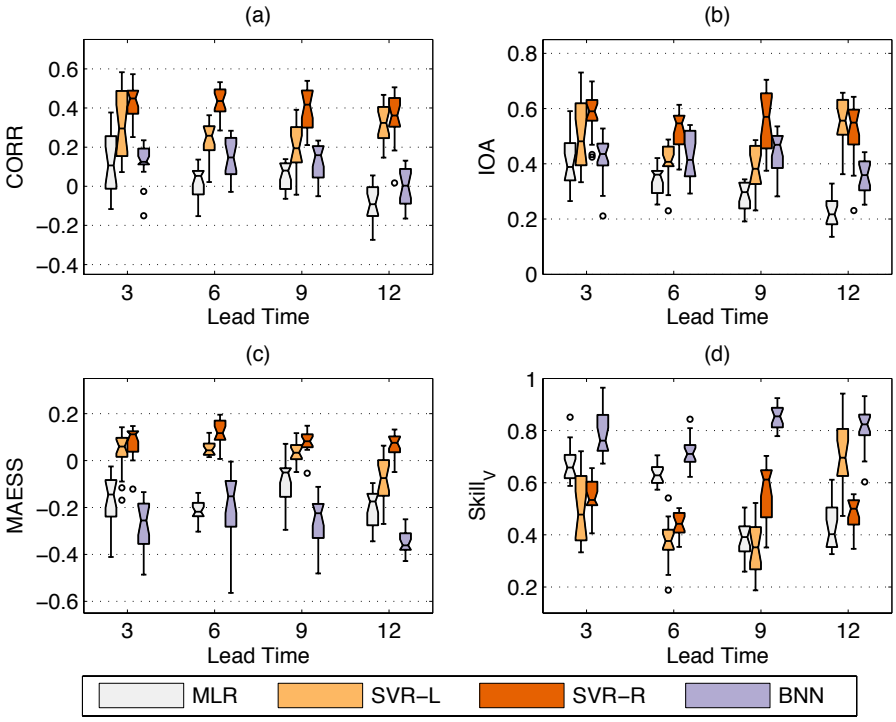

Figure 4: Same as Fig. 2, except over the Prairies (region R3).

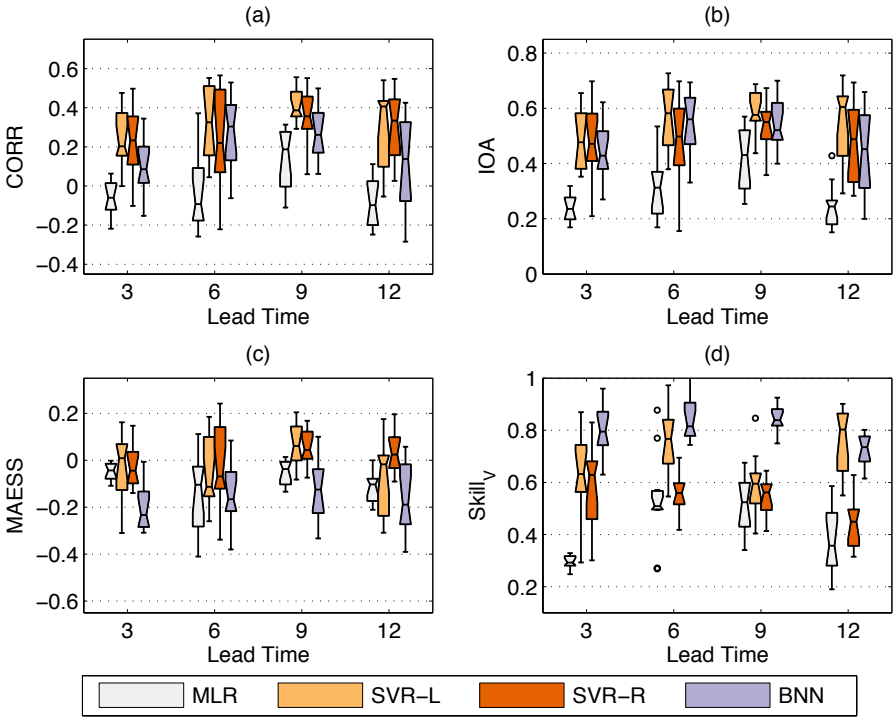

Figure 5: Same as Fig. 2, except over the Arctic (region R4).

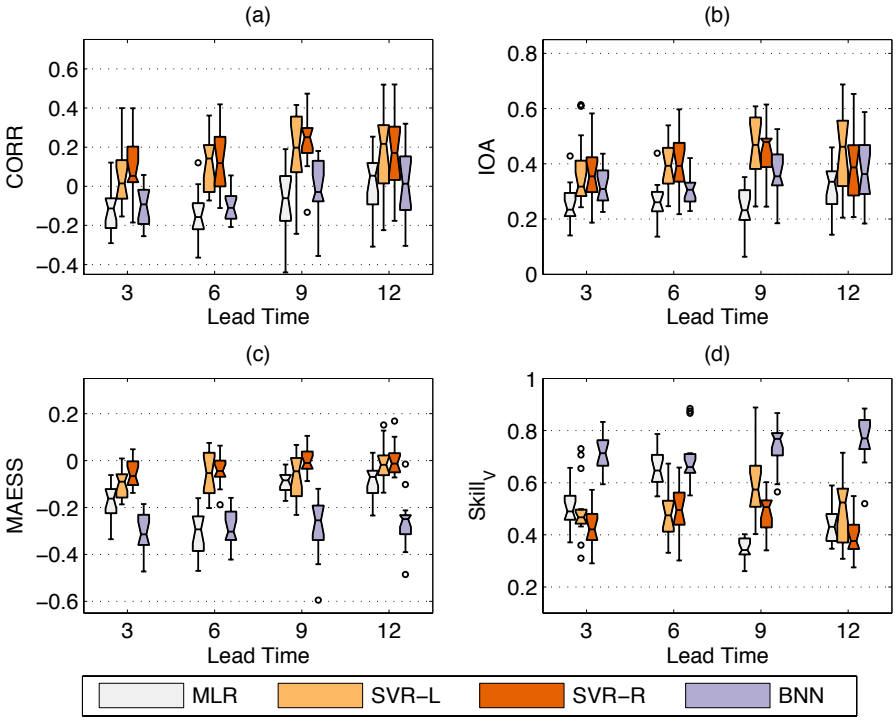

Figure 6: Same as Fig. 2, except over the Great Lakes (region R5).

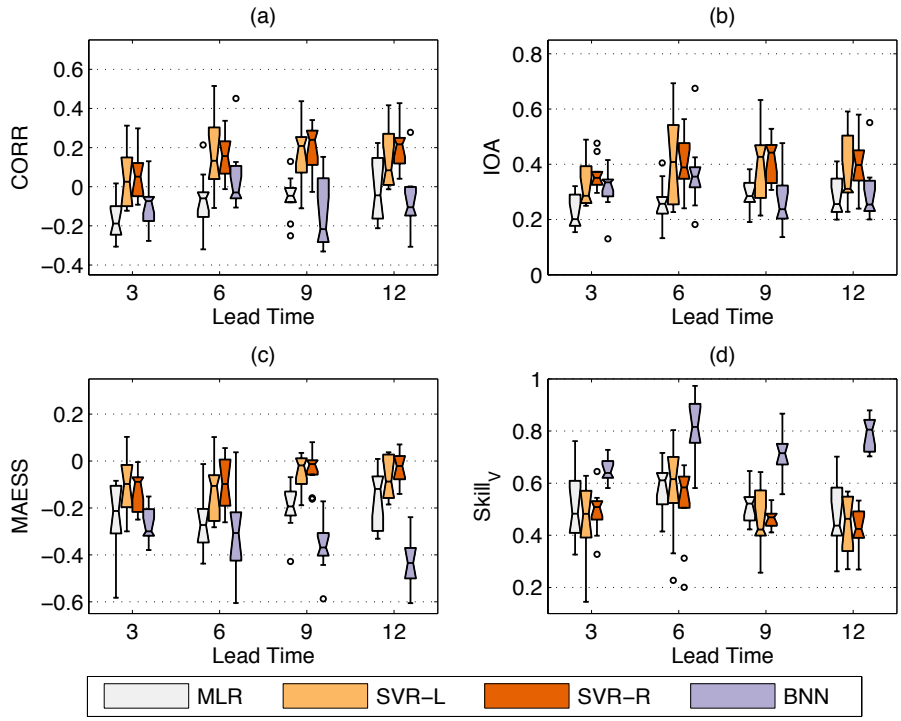

Figure 7: Same as Fig. 2, except over the Atlantic coast (region R6).

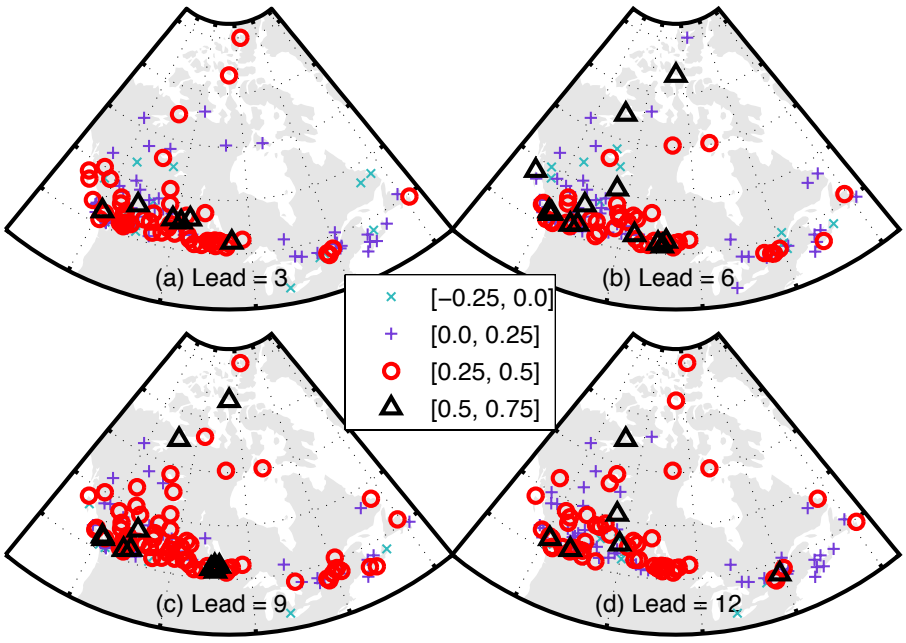

Figure 8: Spatial distribution of the forecast correlation skills of the SVR-R model at individual stations over Canada at lead times of **(a)** 3, **(b)** 6, **(c)** 9 and **(d)** 12 months.

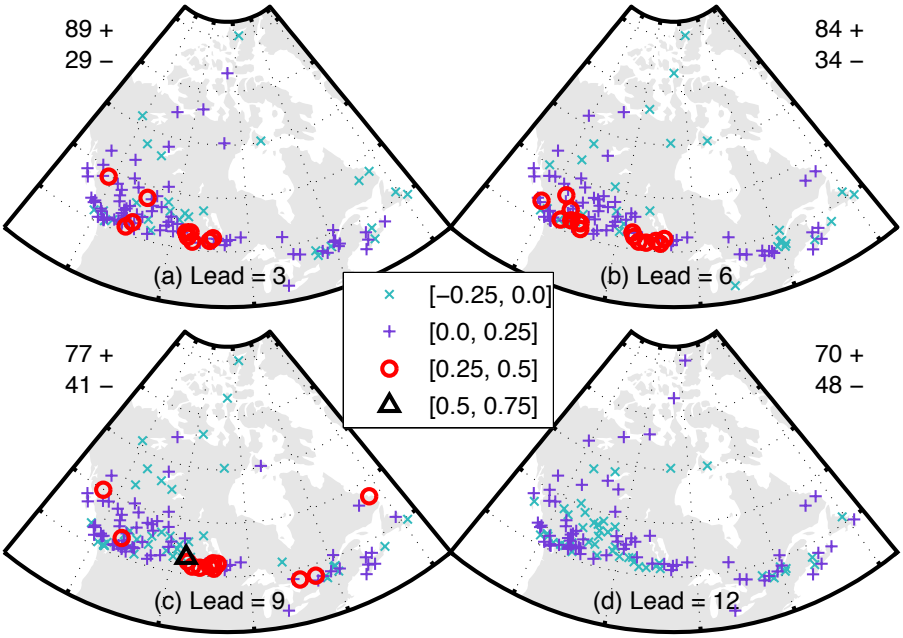

Figure 9: Difference between the forecast correlation skills of the nonlinear SVR model (SVR-R) and that of the linear SVR model (SVR-L) at lead times of **(a)** 3, **(b)** 6, **(c)** 9 and **(d)** 12 months. The two numbers beside each panel give the number of stations where the SVR-R correlation is higher (lower) than that of the SVR-L model, as indicated by the $+(-)$ sign.

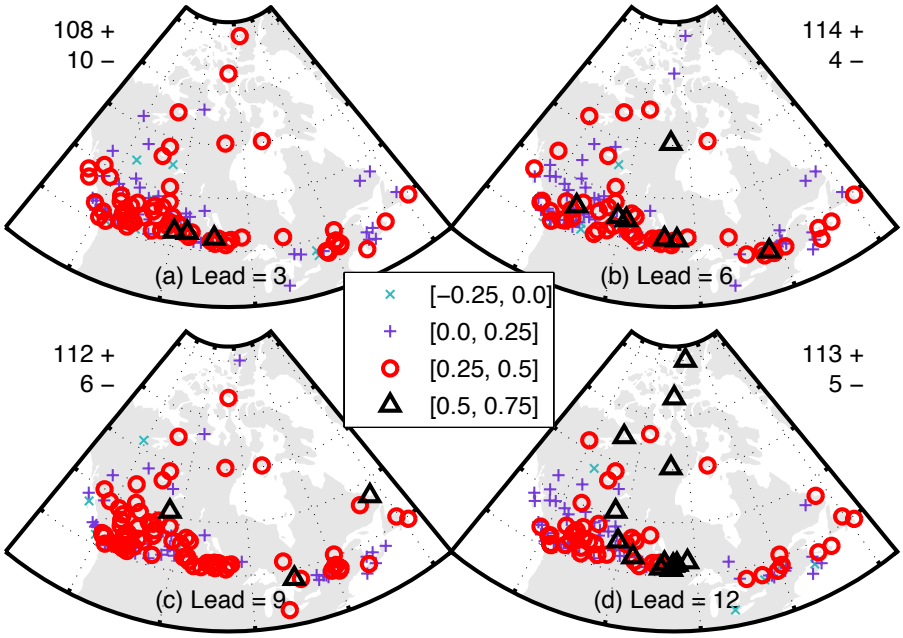

Figure 10: Difference between the forecast correlation skills of the SVR-R model and that of the MLR model at lead times of **(a)** 3, **(b)** 6, **(c)** 9 and **(d)** 12 months. The two numbers beside each panel give the number of stations where the SVR-R correlation is higher (lower) than that of the MLR model, as indicated by the $+(-)$ sign.

15) "Subsection 3.3: The description of double cross-validation procedure lacks clarity. It would be better if equations are given to explain this part of the analysis. (a) Describe in detail the procedure followed to determine optimal number of preditand PCs. (b) Complete results of the double cross-validation must be presented in the manuscript for obvious reasons."

We will add a new Figure 11 to clarify the double cross-validation procedure.

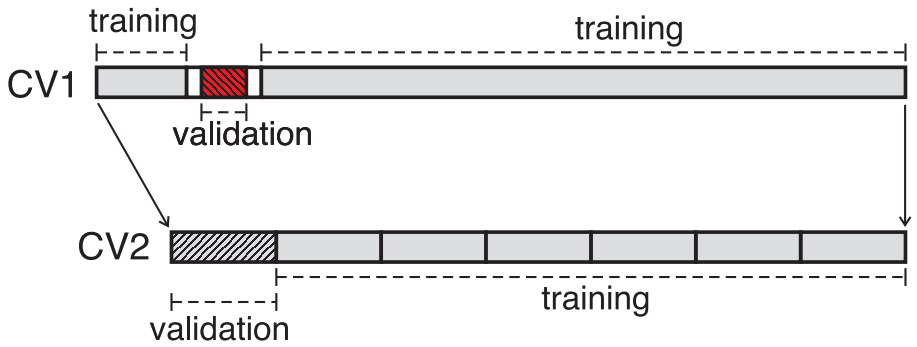

Figure 11: Schematic diagram illustrating the double cross-validation procedure. In the outer round (CV1), the training data are shown in grey and the 3-year validation data shaded. The 1-year data segments (shown in white) bridging the training data and the validation data are not used, to avoid autocorrelation leaking information from the training data to the adjacent validation data. The validation data segment is moved repeatedly in 3-year increments from the start of the data record to the end in this cross-validation loop, so forecast performance was validated over the whole record. Meanwhile in the inner loop (CV2), the training data from CV1 are assembled and divided into 7 segments, with 6 used for training and one (shaded) for validation. Again the training and validation segments are rotated in the loop so all segments are eventually used for validation to determine the optimal model values/hyperparameters. The optimal model determined from CV2 is then used to forecast over the 3-year validation segment in CV1.

16) "Page 3530, lines 18-20: 'For seasonal forecasting, the sample size to the number of predictors is relatively small, since we have 5 SSTPCs, 5 Z500PCs and 6 climate indices as predictors. Hence PCA is again applied to these predictor time series to further reduce the number of predictors.'
(a) What is the maximum number for predictors that can be considered for the sample size being analyzed? Are there any guidelines for deciding the number of predictors for given sample size?
(b) What is the reason for applying PCA to predictors comprising of 10PCs (apart from 6 climate indices)? If 10 PCs were excessive, why were so many PCs extracted? Is it logical to apply PCA to PCs? If the resulting PCs are also excessive do authors suggest applying PCA once more?"

(a) There are no hard rules for maximum number of predictors since regularization

can reduce model complexity via large weight penalty. However, it is generally a good idea to keep the number of predictors less than the number of independent samples. From experience, we have found that reducing excessive predictors to be important for improving forecast skills.
(b) The two PCA serve very different purposes. The first PCA is to do a space-*time* compression of a single field (e.g. SST or Z500), while the second is to compress the relatively large number of input variables (5 PCs from SST, 5 PCs from Z500 and 6 climate indices). There is no need for any further compression.

17) "Page 3531, lines 10-12: 'Forecast testing was only done on the middle 3 yr of the 5-yr data segment to alleviate the leakage of low-frequency signals from the training data to the adjacent test data.
The write-up lacks clarity, and the procedure has to be explained in detail for the sake of readers. What is leakage of low-frequency signals?"

With the new Fig.11 added, this should now be clear. Leakage of low-frequency signals means the autocorrelation of the time series allows low-frequency signals to leak information between the training data and the adjacent validation (testing) data.

18) "Page 3531, lines 20-22: 'For BNN, the optimal number of hidden neurons to use in a neural network model was found from CV2.' The procedure must be explained in detail for obvious reasons."

The cross-validation loop CV2 computes model forecasts skills over validation data for models with different number of hidden neurons and selects the optimal model to use in computing the forecast skills over the validation data in CV1 (Fig.11).

19) "Subsection 3.4: What is the reason for choosing only linear measure for estimating correlation?"

The Pearson correlation is a very standard statistic, so we need to show it or some other reviewer will ask for it.

20) "Page 3533, lines 12-14: 'Ironically, BNN had...' Ironically Skill_V score of SVR-R model is least of all models, which according to author is worst performance. What is the reason for the worst performance?"

Since the models can only explain a very small part of the variance in the data, the model forecasts are expected to have much smaller standard deviation than the data. As the robust SVR-R method is not overfitting (i.e. fitting to noise in the data) like BNN, so its forecasts have smaller standard deviation (hence lower Skill_V) than BNN forecasts, but are more accurate than BNN in terms of MAESS and other measures. In other words, BNN forecasts have larger variance, but the forecasts match more poorly with the observations than SVR – displaying the classic symptom of overfitting.

21) "Page 3533, lines 14-15: All the models are underpredicting standard deviation. Justify using the terminology 'overfitted' for BNN model. How to decide whether a model overfitted data? What can be concluded if Skill_V is nearly zero, less than one, and greater than one?"

See our answer for 20). Skill_V should be less than 1, since the models can only explain a small portion of the variance in the data.

22) Latest references relevant to the study are not referred, whether it is forecasting or cluster analysis (e.g., Partal and Kisi, 2007).

We will update the references in the revised manuscript.

23) "Page 3534, lines 12-13: Why comparison is not presented with results from the canonical correlation analysis prediction model?"

The Shabbar and Barnston (1996) CCA model was for seasonal mean precipitation, not seasonal extreme precipitation.

24) "Section 4: There are 4 skill scores and 4 models. How is it decided whether a region showed highest forecast skill at a particular lead time. Is the judgment purely subjective? The details must be provided for obvious reasons."

We consider the MAESS the most important skill score of the 4. IOA is a combination of the mean squared error and CORR, so is probably the second most important. CORR is widely used, so it is third, while Skill_V is only useful for showing the standard deviation of the model forecasts relative to the observed. There is general agreement between the first 3 in a region, hence our decision is reasonable.

25) "Page 3536, lines 6-7: 'The strongest nonlinearity was found over the Eastern Prairies according to the difference in the forecast performance between the SVR-R and SVR-L models.'
Page 3536, lines 17-18: '. . . we found highest skill in the Eastern Prairies, presumably due to the strong nonlinear signal there,. . . '
Did authors use any standard procedures to detect the presence of nonlinearity in data? Or did they speculate the presence of nonlinearity in data based on the difference in the forecast performance between the linear and nonlinear models?"

The presence of nonlinearity in data was inferred based on the difference in the forecast performance between the linear and nonlinear models.

26) "Page 3536, last paragraph: Even if the contribution to forecast skill (from individual predictors) cannot be quantified quantitatively, is it not possible to arrive at subjective conclusion based on analysis? It is necessary to state which predictor contributes more to forecast skill in each region?"

We do not know of any way to do this.

Minor comments:

Minor comments, mainly of an editorial nature, will be taken care of in the revised manuscript.