A Comparison of Bayesian and Conditional Density Models in Probabilistic Ozone Forecasting

Song Cai, William W. Hsieh, and Alex J. Cannon *Member, INNS*

Abstract— **Probabilistic models were developed to provide predictive distributions of daily maximum surface level ozone concentrations. Five forecast models were compared at two stations (Chilliwack and Surrey) in the Lower Fraser Valley of British Columbia, Canada, with local meteorological variables used as predictors. The models were of two types, conditional density models and Bayesian models. The Bayesian models (especially the Gaussian Processes) gave better forecasts for extreme events, namely poor air quality events defined as having ozone concentration $\geq$ 82 ppb.**

I. Introduction

FORECASTING poor air quality events associated with high surface level ozone concentration has had generally low skills [4]. The motivation behind developing probabilistic forecast models is to provide reliable predictive distribution for these extreme events.

In [4], the procedure is to use statistical-machine learning methods to forecast the daily maximum ozone concentration from local meteorological measurements of the same day plus the ozone concentration from the previous day. After the empirical model has been built, numerical weather prediction model output for the next day is used as input to the empirical model to issue operational ozone forecast for the next day. In this paper, we will compare several probabilistic models on forecasting the daily maximum surface level ozone concentration for two stations (Chilliwack and Surrey) in the Lower Fraser Valley (LFV) of British Columbia, Canada. Of particular interest are poor air quality events, defined as having daily maximum ozone concentration $\geq$ 82 ppb.

We will compare two main types of models capable of forecasting distributions: Conditional density models [1], and Bayesian models [2], [11]. The main difference between the two is that the former uses a single optimal function (based on maximum likelihood) to forecast while the latter gives all probable functions a non-zero probability and integrates over all of them to obtain the forecast. We will show that by including low probability functions, the Bayesian approach forecasts extreme events better.

II. Methodology

A. Data

Data were obtained from the Greater Vancouver Regional District (GVRD) air quality monitoring network [4]. As the majority of elevated ozone concentrations in the region is confined to late spring and summer, 947 days of data were extracted for the months from May through September, 1994–2001. The single predictand or response variable is the daily maximum of the hourly surface level ozone concentration at one station. We will test predictions at two separate stations, Chilliwack and Surrey in the Lower Fraser Valley of British Columbia. Since ozone concentration is non-negative, we used the natural logarithm of the ozone concentration for the predictand, so the predictand is an unbounded real variable.

There are 29 potential predictors, these being the maximum daily ozone concentration of the previous day, and the temperature, precipitation, pressure, wind speed and direction (all at zero lead time) from several stations in the monitoring network (similar to [4]). Random Forests [3], an ensemble version of Classification and Regression Trees (CART), was used to reduce the number of predictors, from 29 to 16 for Chilliwack, and to 20 for Surrey.

B. Models

Five models are used. Two are conditional density models: Conditional density network with Gaussian distribution (CDN-Gaussian) [1], [14], [5], and conditional density network with the Johnson translation system (CDN-Johnson) [12]. Both CDN models use multi-layer perceptron (MLP) neural network models (with 1 hidden layer) to model the parameters of the distributions, with the cost function being the negative log likelihood, and overfitting prevented by early stopping [1]. In CDN-Gaussian, the predictand (i.e. the log of the ozone concentration) is assumed to have a Gaussian distribution. Once we get the predictive distribution of the log ozone concentration, we can easily calculate the predictive distribution of the ozone concentration, which is a log-normal distribution. From the Johnson translation system of distributions, the 3-parameter log-normal distribution was chosen, with the 3 parameters modeled by an MLP. Since this Johnson distribution is log-normal, the ozone concentration is used directly as the predictand in CDN-Johnson.

The remaining three models use the Bayesian framework. For all three models, we take the log of the ozone concentration as the predictand and assume that the predictand has a Gaussian distribution, as we did for CDN-Gaussian. The predictive distribution of ozone concentration is therefore

S. Cai is with the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (scai@eos.ubc.ca).

W. W. Hsieh is with the Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (phone: 604-822-2821; fax: 604-822-6088; email: whsieh@eos.ubc.ca).

A. J. Cannon is with the Meteorological Service of Canada, Environment Canada, 201-401 Burrard Street, Vancouver, BC V6C 3S5, Canada (alex.cannon@ec.gc.ca).

log-normal for these models. In Gaussian process (GP) regression [11], an input-independent Gaussian noise and a Gaussian process prior are used. Being a kernel method, GP offers a choice of kernel or covariance functions. In this case, a commonly used Matern class kernel function [11] with parameter 5/2 is used. Bayesian neural network (BNN) [9] also uses an input-independent Gaussian noise and a Gaussian prior and is trained with a 1-hidden-layer MLP. Finally, Bayesian linear regression (GP-linear) is also used to check whether the underlying relations are linear or nonlinear. This is not a least-square linear regression, but a Bayesian version of linear regression, done by restricting the covariance function to a linear function in GP.

Our particular interest is on GP, a relatively new method, and one purpose of this paper is to compare it with neural networks. Using a Bayesian framework, GP has the all the advantages of Bayesian techniques in machine learning, such as naturally avoiding overfitting. Moreover, with the structure as specifed in our paper, GP is analytical tractable, so its accuracy is high relative to other Bayesian learning methods which need to make approximations when calculating the posterior distribution. Being a kernel method, GP can model a broad range of non-linear functions by using suitable kernels, hence it can capture complicated non-linear relationships just as neural networks. One criticism of GP is that it is computationally expensive. In practice, we found that the speed of GP and neural networks are comparable — the reason being that neural networks need to average over an ensemble of models to alleviate local minima in the cost function.

An 8-fold cross-validation scheme was used, i.e. with 8 years of data, 1 year was set aside for forecast verification, while the models were trained using data from the remaining 7 years, and the process was repeated until all 8 years have been used for forecast verification. For all the models using neural networks (CDN-Gaussian, CDN-Johnson and BNN), in each fold of cross-validation, the bootstrap procedure [6] was used to randomly select bootstrap samples from the training data. A model was trained for each bootstrap sample – the training data not selected by the particular bootstrap sample would be used later for testing the trained model. For CDN-Gaussian and BNN, the results were taken as the ensemble average of 300 bootstrap samples, and for CDN-Johnson, the average of 15. The bootstrap ensembles were repeated for different number of hidden neurons, and from the testing data, the optimal number of hidden neurons was determined.

III. Results

A. Deterministic Scores

The median of the predictive distribution can be used to give a deterministic forecast. The mean absolute error (MAE) and root mean square error (RMSE) at Chilliwack and Surrey are shown in Table I. The results show that for both stations and in terms of both MAE and RMSE, GP turned out to be the best model. But these results are not very meaningful for probabilistic forecasting models, since the decision probability threshold does not necessarily have to be the median, especially for extreme weather events.

TABLE I
Deterministic scores for the five models, with the best score among the five models printed in bold.

	Chilliwack		Surrey	
	MAE	RMSE	MAE	RMSE
GP	**5.3534**	**7.0045**	**5.0334**	**6.5849**
BNN	5.7443	7.8145	5.4675	7.3937
CDN-Gaussian	5.9034	7.7082	5.5122	7.1279
CDN-Johnson	5.5074	7.2042	5.4016	7.0437
GP-Linear	6.2890	8.2665	5.8283	7.5295

B. Probabilistic Scores

The forecasted and observed ozone concentration at Chilliwack are plotted for 120 days during May-Sept., 1995 in Fig. 1 (with data gaps omitted) for the GP and CDN-Johnson results. For the first 20-30 points and the most extreme value, the 95% prediction interval for GP is wider than that for CDN-Johnson, but for the remaining part, the prediction interval for GP is comparable to that for CDN-Johnson.

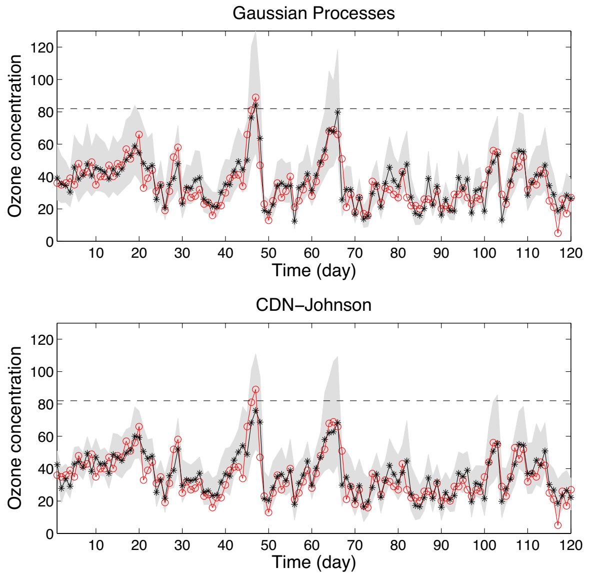

Fig. 1. Forecast of daily maximum ozone concentration at Chilliwack by the GP model (upper panel) and by the CDN-Johnson model (lower panel). The ozone concentration is plotted, with the threshold for poor air quality events (i.e. ozone concentration $\geq$ 82 ppb) indicated by the horizontal dashed line. The median of the predictive distribution for each day is indicated by an asterisk, with the observed value indicated by a circle. The 95% prediction interval is shaded.

A good probabilistic forecast should have two attributes: reliability and sharpness [10]. Reliability means that the predictive probability of an event should be consistent with the historical observations, and sharpness means that the predictive probability should separate from the climatological probability forecast. Well-designed scores for evaluating probabilistic forecasts of continuous variables are the continuous ranked probability score (CRPS) and the ignorance score (IGN) [7].

The continuous ranked probability score is defined as

$$\begin{aligned} \mathrm{CRPS} &= \frac{1}{n}\sum_{i=1}^{n}\mathrm{crps}\left(F_i,\, y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\int_{-\infty}^{\infty}\left[F_i\left(y\right)-H\left(y-y_i\right)\right]^2\mathrm{d}y\right), \end{aligned}$$

where for the ith prediction, the cumulative probability $F_i\left(y\right)=p\left(Y\le y\right)$, and $H\left(y-y_i\right)$ is the Heaviside function that takes the value 0 when $y-y_i<0$, and 1 otherwise.

The ignorance score is defined as

$$\mathrm{IGN}=\frac{1}{n}\sum_{i=1}^{n}\mathrm{ign}\left(p_i,\, y_i\right)=\frac{1}{n}\sum_{i=1}^{n}\left[-\log\left(p_i\left(y_i\right)\right)\right],$$

where p_i is the predictive density and y_i the corresponding observed value. IGN is simply the negative log predictive density, which is also the cost function used to train the CDN models.

Both scores are negatively oriented, i.e. the lower the better. If the predictive distribution is Gaussian (with mean μ and standard deviation σ), the analytical forms of the scores can be derived [7]. For a Gaussian distribution, the key difference between these two scores is that CRPS grows linearly with the normalized prediction error $\left(y-\mu\right)/\sigma$, but IGN grows quadratically. Hence, the ignorance score assigns harsh penalties to particularly poor probabilistic forecasts, and can be exceedingly sensitive to outliers and extreme events [13], [8].

The CRPS and IGN averaged over all the test points are shown in Table II, the scores over the poor air quality events, in Table III, and the scores over the fair air quality events (i.e. 52 ppb $\le$ ozone concentration $<$ 82 ppb), in Table IV. If we care only about the poor air quality and treat it as a binary event, then we can also calculate the Brier score (BS), Brier skill score (BSS) and the area under the Relative Operating Characteristic (ROC) curve [10]. However, there are only 9 poor events out of 947 points for Chilliwick and 5 out of 947 points for Surrey, so when calculating those scores, the calculation error may be larger than the differences between the scores of different models, which makes them unmeaningful. If we take 52 ppb as the threshold of the binary forecast (i.e. fair-poor air quality versus good air quality), the sample size should be enough to calculate reliable BS, BSS and area under the ROC curve. These scores are shown in Table V, with BS being negatively oriented, BSS and area under the ROC curve being positively oriented, and climatology used as the reference forecast in calculating BSS.

For the overall scores (averaged over all 947 test points), among the four top scores printed in bold in Table II, GP captured three of the four, while CDN-Johnson captured one. Further behind are CDN-Gaussian, BNN and GP-linear.

TABLE II

CRPS AND IGN CALCULATED OVER ALL 947 TEST POINTS, WITH THE BEST SCORE AMONG THE FIVE MODELS PRINTED IN BOLD.

	Chilliwack		Surrey	
	CRPS	IGN	CRPS	IGN
GP	**3.8708**	3.4105	**3.6342**	**3.3268**
BNN	4.2610	3.4181	3.9591	3.3535
CDN-Gaussian	4.2148	3.9990	3.9159	3.3570
CDN-Johnson	3.9083	**3.3575**	3.8444	3.3652
GP-Linear	4.5292	3.5508	4.1604	3.4658

TABLE III

CRPS AND IGN CALCULATED OVER POOR AIR QUALITY EVENTS. FOR CHILLIWACK, THERE ARE 9 POOR EVENTS OUT OF 947 POINTS; FOR SURREY, THERE ARE 5 POOR EVENTS OUT OF 947.

	Chilliwack		Surrey	
	CRPS	IGN	CRPS	IGN
GP	**8.6306**	**4.2607**	**13.3502**	**4.7924**
BNN	11.7326	4.3519	16.0328	5.0461
CDN-Gaussian	10.2731	4.5850	16.5075	5.4449
CDN-Johnson	11.7009	4.5932	16.9548	5.2019
GP-Linear	10.0993	4.4348	16.5386	5.2284

For the scores averaged over only the poor air quality events (Table III), the differences between Bayesian models and conditional density models are remarkable. Among the four top scores printed in bold, GP captured all four. Although BNN performed mediocrely on overall scores, it captured three second-best scores for poor events. Even GP-linear, a linear Bayesian model performed better than the two nonlinear conditional density models for Chilliwack and comparable to them for Surrey over poor events.

The 9 poor events at Chilliwack plotted in Fig. 2 revealed that GP performed better than CDN-Johnson for most poor events. The upper panel shows the predictive median and

TABLE IV

CRPS AND IGN CALCULATED OVER FAIR AIR QUALITY EVENTS. FOR CHILLIWACK, THERE ARE 110 FAIR EVENTS OUT OF 947 POINTS; FOR SURREY, THERE ARE 114 FAIR EVENTS OUT OF 947.

	Chilliwack		Surrey	
	CRPS	IGN	CRPS	IGN
GP	5.9475	3.8493	**5.4853**	**3.7728**
BNN	5.9389	**3.8315**	6.2033	3.8775
CDN-Gaussian	6.6425	4.1880	6.1155	3.8815
CDN-Johnson	**5.9080**	3.8386	5.8706	3.8611
GP-Linear	7.2645	4.0657	6.5405	3.9429

TABLE V

BS, BSS AND THE AREA UNDER THE ROC CURVE CALCULATED USING THE THRESHOLD OF 52 PPB FOR BINARY FORECASTS AT CHILLIWACK AND SURREY.

	Chilliwack			Surrey		
	BS	BSS	ROC	BS	BSS	ROC
GP	**0.0519**	**0.527**	0.957	**0.0621**	**0.434**	**0.934**
BNN	0.0525	0.522	0.957	0.0664	0.396	0.924
CDN-Gauss.	0.0570	0.482	0.946	0.0679	0.382	0.923
CDN-Johnson	**0.0519**	**0.527**	**0.964**	0.0677	0.384	0.929
GP-Linear	0.0644	0.414	0.935	0.0740	0.327	0.909

the 95% prediction interval for GP and for CDN-Johnson. Both predictive medians of GP and CDN-Johnson underpredicted the ozone concentration for all 9 events. For the 1st event, CDN-Johnson outperformed GP, but for the 2nd to 7th events, GP outperformed CDN-Johnson, while for the last two events, GP and CDN-Johnson were similar. The 95% prediction intervals for GP and CDN-Johnson largely overlapped, with both encompassing the true values. The lower panel also shows the predictive distributions as a function of the ozone concentration for the 1st, 4th and 7th events. For the 4th and 7th events, the predictive distributions of GP are skewed towards larger ozone concentration values than those of CDN-Johnson; for the first event, the situation is reversed.

For the scores over fair events (Table IV), CDN-Johnson, BNN and GP are comparable for Chilliwack, while for Surrey, GP has the highest scores and CDN-Johnson second highest. Overall these differences between Baysian nonlinear models and conditional density models are small, i.e. we can consider them comparable over fair events. These results agree with the BS, BSS and area under ROC curve for binary forecast with threshold at 52 ppb (Table V).

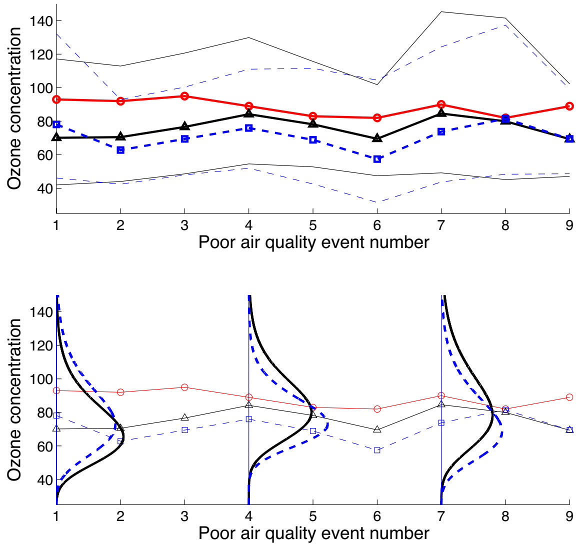

Fig. 2. Prediction and predictive distribution of the ozone concentration for the 9 poor air quality events at Chilliwack. The predictive medians are indicated by triangles connected by a solid line in GP and squares connected by a dashed line in CDN-Johnson, with the corresponding observed values shown as circles. The thin solid and dashed lines in the top panel indicate the 95% prediction interval for GP and CDN-Johnson, respectively. Bottom panel also plots the predictive distributions for the 1st, 4th and 7th events, with the solid curve for GP and dashed curve for CDN-Johnson, and with the predictive medians and observations in the upper panel reproduced as thin lines.

IV. Discussion and Conclusion

In theory, Bayesian models can give an accurate measure of the predictive uncertainty arising from (a) the uncertainty of the noise process and (b) the uncertainty of the model weights (i.e. parameters) due to finite sample size. The conditional density models only estimates the predictive uncertainty arising from the noise process, without taking in account the uncertainty of the model weights.

In the observed data, most of the test points have good similarities with the training data, so for these points, the uncertainty of the model weights (or uncertainty of the underlying function) is low. Using the weights found by maximizing likelihood, conditional density models tend to find a function which is quite close to the true underlying function. Consequently, it will give good prediction for these test points. On the other hand, Bayesian approaches give the most probable functions a very high probability, but it does not rule out other possibilities — instead, it gives the unlikely functions a very low but nonzero probability. Therefore, for low uncertainty points, Bayesian models can have comparable (or slightly worse) performance relative to conditional density models. For the relatively few points which have little similarity with the training data, the uncertainty of the underlying function (hence the model weights) is high. Conditional density models just decide on one function and rule out other functions, while Bayesian models give all possible functions a non-zero probability, and integrate over all of them to obtain the forecast. Thus in general, the Bayesian models have better performance over the highly uncertain events. This is the reason why Bayesian models may have similar overall scores compared to conditional density models, but outperform them over the rare events.

The conditional density models can be tuned to achieve good performance on rare events: For the CDN models, we can increase the number of hidden neurons and do not use regularization (or decrease the number of hidden neurons if the rare events have low variance), thus forcing the model to fit the few rare events. Alternatively, when re-sampling, we can increase the number of repetitions for the rare events [4], which is equivalent to converting the rare events to "normal" events. But applying these techniques will sacrifice the performance of the model on the majority of the data points and will yield worse overall scores, hence a trade-off.

Using a Bayesian approach, GP provides a moderate solution for this problem, i.e. it gives good overall scores and also good scores on rare events, making GP particularly valuable in extreme weather forecasting. The relatively mediocre performance of the BNN is not entirely clear. It is probably because the BNN code (from Netlab [9]) makes the Laplace approximation when solving for the hyperparameters. In contrast, the GP Bayesian formulation is analytically tractable without having to make the Laplace approximation — an attractive feature of GP relative to other Bayesian methods.

Acknowledgment

This work is supported by the Canadian Foundation for Climate and Atmospheric Sciences, and the Natural Sciences and Engineeering Research Council of Canada.

References

[1] C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford: Clarendon, 1995.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, New York: Springer, 2006.

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[4] A. J. Cannon and E. R. Lord, "Forecasting summertime surface-level ozone concentrations in the Lower Fraser Valley of British Columbia: An ensemble neural network approach," *J. Air & Waste Manage. Assoc.*, vol. 50, pp. 322–339, 2000.

[5] S. R. Dorling, R. J. Foxall, D. P. Mandic and G. C. Cawley, "Maximum likelihood cost functions for neural network models of air quality data," *Atmos. Environ.*, vol. 37, pp. 3435–3443, 2003.

[6] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *J. Amer. Stat. Assoc.*, vol. 92, pp. 548–560, 1997.

[7] T. Gneiting, A. E. Raftery, A. H. Westveld III and T. Goldman, "Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation," *Mon. Wea. Rev.*, vol. 133, pp. 1098–1118, 2005.

[8] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Amer. Stat. Assoc.*, vol. 102, pp. 359–378, 2007.

[9] I. T. Nabney, *Netlab: Algorithms for Pattern Recognition.*, London: Springer, 2002.

[10] I. T. Jolliffe and D. B. Stephenson, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, Chichester: Wiley, 2003.

[11] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT, 2005.

[12] W. C. Torrez and J. T. Durham, "Johnson distribution for fitting weighted sums of sigmoided neuron outputs," *Technical report, Signal and Information Processing Division, NCCOSC RDT&E Division, 53560 Hull Street, San Diego, CA 92512-5001*, 1993.

[13] A. S. Weigend and S. Shi, "Predicting daily probability distributions of S&P 500 returns," *J. Forecasting*, vol. 19, pp. 375–392, 2000.

[14] P. M. Williams, "Modelling seasonality and trends in daily rainfall data," in *NIPS '97: Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*, vol. 10, pp. 985–991, 1998.