Modeling Weather-induced Home Insurance Risks with Support Vector Machine Regression*

Asim Kumer Dey†   Vyacheslav Lyubchich‡   Yulia R. Gel*

Abstract

Insurance industry is one of the most vulnerable sectors to climate change. Assessment of future number of claims and incurred losses is critical for disaster preparedness and risk management. In this project, we study the effect of precipitation on a joint dynamics of weather-induced home insurance claims and losses. We discuss utility and limitations of such machine learning procedures as Support Vector Machines and Artificial Neural Networks, in forecasting future claim dynamics and evaluating associated uncertainties. We illustrate our approach by application to attribution analysis and forecasting of weather-induced home insurance claims in a middle-sized city in the Canadian Prairies.

1 Introduction.

Changes in extreme weather and climate events, such as heat waves, droughts, heavy rainfall, and flood are the primary way that people experience climate change. Over the last 50 years, much of the United States has seen increases in prolonged periods of excessively high temperatures, droughts, and, in some regions, high precipitation and severe floods. Since 1950, extreme precipitation events have become more common in many regions in the United States [20]. Heavy precipitation is the primary cause of floods, when streams and rivers overflow their banks. Since 2008, the United States has seen six floods costing at least $1 billion each, resulting in damaged property and infrastructure, agricultural losses, displaced families, and even loss of life [4], [25]. Similarly to the USA, the number of flood disasters in Canada has noticeably increased in the past decades [21], [3], [4].

Within the financial sector, the insurance industry faces the most substantial challenges from such climatic changes. From 2009 to 2014, in Canada, total insured losses from catastrophic events were close to or above $1 billion each year—most caused by water damage. In fact, in 2013 Canadian insurers paid out a record-high $3.2 billion to policy holders [14]. The total rainfall related water damage incurred losses for the period of April–September in 1992–2002 sum up to about $970 million for the four selected Canadian cities combined (London, Kitchener-Waterloo, Toronto, and Ottawa) [8]. The Intergovernmental Panel on Climate Change (IPCC) has projected that the severity and frequency of extreme rainfalls (and, consequently, floods) will further increase [16]. Thus, the flood related losses will rise in the future, jeopardizing financial stability of the insurance industry and the whole society [25].

Various conventional statistical approaches, e.g., Generalized Linear Model (GLM), Autoregressive Integrated Moving Average (ARIMA) model, and Bayesian hierarchical model, have been employed for modeling the number of insurance claims (see the recent reviews by [12], [26], [19], [22] and references therein). More recently, different machine learning techniques, such as Neural Network (NN), Support Vector Machine (SVM), and Random Forest, have also been successfully utilized for assessing dynamics of claim frequencies [30], [5]. Nevertheless, there exists a gap in understanding on how weather-induced losses, or dollar magnitudes of claims, depend on excessive precipitation and how to *simultaneously* model and forecast a joint dynamics of weather-related losses and frequencies.

We propose a new two-stage procedure to bridge the above gap based on Support Vector Regression (SVR) with Genetic Algorithm (GA), further referred to as GA-SVR. The main contributions of this paper are as follows:

- We propose a new parsimonious two-stage approach for the *joint* attribution analysis of weather-induced home insurance claims (frequencies) and losses (severities). We show that the new nonparametric attribution procedure delivers a competitive performance in characterizing spikes in the time series of claims and losses, while using a single exogenous regressor, i.e., precipitation.

- We use a computationally efficient data-driven GA

approach to optimize hyper-parameters of SVR. We evaluate the new SVR-GA procedure for weather-induced risk attribution in respect to Artificial Neural Networks (ANN) and conventional SVR procedures. We find that SVR-GA leads to substantial gains in performance, that is, SVR-GA delivers at least 70% lower root mean squared error for the number of claims and more than 10 times lower root mean squared error for aggregate losses than the competing approaches.

- We develop projected climate-induced numbers of claims and losses for 2021–2080 and investigate a relative change in expected claim dynamics vs. the insurance claims observed over a baseline period of 2002–2011. The forecasts are governed by the latest climate projections delivered by the Canadian Earth System Model (CanESM2) and downscaled by the Canadian Regional Climate Model (CanRCM4).

The remainder of this paper is organized as follows. In §2 we describe insurance and weather data employed in our analysis. The GA-SVR technique is discussed in §3. The results of modeling and forecasting are reported in §4 and discussed in §5.

2 Data.

The data comprise of daily rainfall related home insurance claims and losses recorded in a middle-sized city in the Canadian Prairies (City A) in the 10-year period of 2002–2011. The city name is suppressed due to data confidentiality. To remove the effects of changing risk exposure, the number of claims is normalized by the number of homes insured in City A on each day from 2002 to 2011. To remove the effect of inflation, home insurance losses are converted to the prices of 2002 (CAD2002) using metropolitan area composite index of apartment building construction. Future daily precipitation data for 2021–2080 are obtained from the CanESM2 global climate model downscaled to 0.22° horizontal grid resolution (approximately 25 km) by CanRCM4 model. Particularly, we use the data from the CanRCM4 grid cell which center is the closest to City A. As the final pre-processing step, all daily data are aggregated to the weekly level.

Two scenarios of Representative Concentration Pathways (RCPs) available from CanRCM4 projections are RCP 4.5 and RCP 8.5. RCPs define possible climate futures depending on the amount and timing of greenhouse gases emissions in the future. RCP 4.5 assumes that the emissions will peak around 2040, then decline. Under the RCP 8.5 scenario, emissions continue to rise throughout the period 2000–2100 [15]. Projected pre-

Table 1: Overview of the data sets

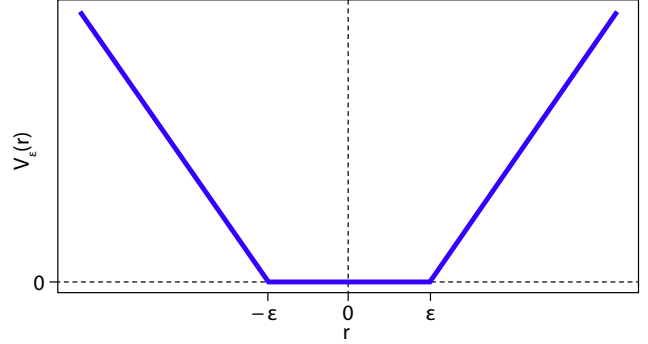| Period	Data type	Variables
Control (2002–2011)	Observations	precipitation, number of claims, losses
Scenario (2021–2080)	Projections RCP 4.5 Projections RCP 8.5	precipitation precipitation

Figure 1: SVR ε-insensitive error function.

cipitation amounts corresponding to these scenarios are used to predict future number of claims as well as aggregate loss. Table 1 provides an overview of the data sets that are used in this study.

3 Support Vector Regression Approach.

Support Vector Regression (SVR) has emerged as an alternative and highly effective tool for solving nonlinear regression and time series problems. Due to its wide applicability, SVR became popular in a variety of forecasting applications, including forecasts of warranty claims [30], volatility of financial returns [7], water requirements [17], and tourism demand [6]. Yuan [31] uses SVR with genetic algorithm to forecast sales volume.

SVR, a version of Support Vector Machine (SVM) [29], is based on the statistical learning theory [28]. In contrast to Least Squares Regression, where in-sample residual sum of squares is minimized, SVR attempts to minimize the generalized error bound. This bound is the combination of the training error and a regularization term that controls the complexity of the hypothesis space [1]. Another important feature of SVR is that the method only depends on a particular subset of the training data, because only residuals that are larger than an absolute value of some constant (ε) contribute to the loss function. For this reason, the SVR loss function is called ε-insensitive loss function (Figure 1).

Let $D = \{(x_i, y_i)\}$ be a learning sample set, where

$x_i \in \mathcal{R}^m$ represents input values; $y_i \in \mathcal{R}$ are corresponding output values for $i = 1, 2, \ldots, N$; N is sample size, and m is dimension of the input dataset. The aim is to identify a regression function, $y = f(x)$, that accurately predicts the outputs $\{y_i\}$ corresponding to a new data set D. In the case of a linear SVR, $f(\cdot)$ has the form

$$(3.1) \qquad f(x) = \langle \beta, x \rangle + \beta_0,$$

where $\beta \in \mathcal{R}^m$, $\beta_0 \in \mathcal{R}$, and $\langle \cdot, \cdot \rangle$ denotes inner product in R^m. To estimate β and β_0, we consider minimization of the following risk function:

$$(3.2) \qquad H(\beta, \beta_0) = \sum_{i=1}^{n} V_\varepsilon \left(y_i - f(x_i) \right) + \frac{\lambda}{2} \|\beta\|^2,$$

where $V_\varepsilon(r)$ is ε-insensitive error function (Figure 1), $\lambda > 0$ is regularization constant, and $\| \cdot \|^2$ denotes Euclidean norm [11], [9].

The risk function, $H(\beta, \beta_0)$, involves summation of the empirical risk (training errors) and regularization term $\|\beta\|^2/2$. The regularization parameter, λ, determines the trade-off between minimizing empirical risk and minimizing the regularization term. Minimization of function (3.2) leads to a quadratic programming problem, which can be converted to the dual Lagrangian problem [11], [6]. Its solution function has the form [1]:

$$(3.3) \qquad \hat{f}(x) = \sum_{i=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i) \langle x, x_i \rangle + \beta_0,$$

where $\hat{\alpha}_i^*$ and $\hat{\alpha}_i$ are positive.

The linear SVR (3.1) can be extended to nonlinear regression through a mapping function $\phi(x)$. We can map every input data point into a high-dimensional feature space through the nonlinear mapping function, $\phi(x)$, and then apply the standard SVR algorithm:

$$(3.4) \qquad \hat{f}(x) = \sum_{i=1}^{n} (\hat{\alpha}_i^* - \hat{\alpha}_i) \langle \phi(x), \phi(x_i) \rangle + \beta_0.$$

The inner product in (3.4) can be replaced by a kernel function $k(x_i, x)$. The kernel function can handle any dimension feature space [27]. Different types of kernel functions are used in SVR. In this study, we use the Radial Basis Function (RBF) kernel (Gaussian kernel):

$$(3.5) \qquad k(x_i, x) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2} \right),$$

where σ^2 is the bandwidth.

However, results of SVR modeling highly depend on a number of user-defined parameters (hyperparameters), which should be carefully selected. Such parameters include:
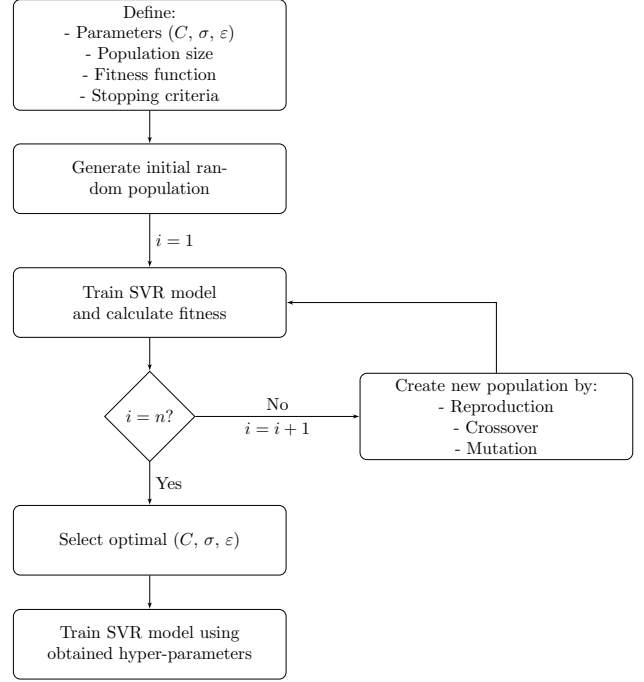

Figure 2: Flow-chart of the GA-SVR.

- Regularization parameter (λ),

- Tube size of ε-insensitive loss function (ε),

- Bandwidth of the kernel function (σ^2).

Inappropriate choice of SVR parameters may lead to over-fitting or under-fitting [18]. An unconnected, separate optimization is not sufficient for finding the optima because these variables influence each other. Different methods can be employed for selecting hyperparameters (see [6] for an overview). In this study, Genetic Algorithm (GA) [10], [24] is applied to simultaneously optimize all SVR parameters. GA is a global search procedure: it checks all solution space from many points for the best combination of free parameters. A general procedure of GA for selecting the optimal hyperparameters is shown in Figure 2. Genetic algorithm uses selection, crossover, and mutation operations to generate the offspring of the existing population [23], [31]. When convergence criteria are met, the algorithm (evolution) stops and delivers the hyper-parameter values corresponding to the best fitness function value.

4 Modeling Precipitation Related Risk.

In our study, the number of claims and aggregate loss are modeled in a two-stage procedure, using Genetic Algorithm Support Vector Regression (GA-SVR). The mapping function for GA-SVR for the number of claims

Table 2: GA parameter settings

Parameter		Value
Number of generations (stopping criterion):		100
Population size:		50
Fitness function:		RMSE
Search	C	$(10^{-3}, 10^3)$
domain [13]:	σ^2	$(10^{-3}, 2^4)$
	ϵ	$(10^{-2}, 2^3)$

is

$$(4.6) \qquad N_t = f(R_t, R_{t-1}, \max R_t)$$

and for the aggregate loss

$$(4.7) \qquad L_t = f(R_t, R_{t-1}, \max R_t, N_t),$$

where N_t is number of claims at week t, L_t is aggregate loss at week t, R_t is total precipitation at week t, and $\max R_t$ is maximum daily total precipitation at week t. GA parameter settings used in this study are given in Table 2.

4.1 Prediction. We fit models (4.6) and (4.7) using the data from the control period (see Table 1) and obtain expected values of weekly number of claims ($\hat{N}_t$) and weekly aggregate loss ($\hat{L}_t$) using the RCP 4.5 and RCP 8.5 data for the scenario period. As the aggregate loss model relies on the number of claims N_t, we start from forecasting future number of claims. Due to the large uncertainties of climate projections at such high—weekly—temporal resolution (see [19] on the potential sources of uncertainties), we consider aggregated trends in the future, rather than individual weekly number of claims or aggregate loss. We summarize the trends as the relative change (measured in percents) from the control period to the projection (scenario) period. For balanced comparisons of scenario with control, we split the 60-year scenario period (2021–2080) into six 10-year sub-periods to make the control period (2002–2011) and scenario periods (2021–2030, 2031–2040, …, 2071–2080) of equal length. Then, we calculate the percentage change in the number of claims or aggregate loss for each of the six scenario sub-periods:

$$(4.8) \qquad \Delta = \left(\frac{\sum_{t \in scn} \hat{N}_t}{\sum_{t \in ctr} N_t} - 1 \right) 100\%,$$

where *scn* and *ctr* refers to the scenario sub-period and control period, respectively. Since both periods are of 10 years, the numerator and denominator are of comparable size, so no further normalization is necessary.

We now assess the expected claim and aggregate loss dynamics corresponding to the two climate change scenarios for the area of City A. Figure 3(a) shows that the annual number of home insurance claims will increase noticeably in the scenario period, relative to the baseline of 2002–2011. The projected number of claims for the climate scenario RCP 4.5 is from 4.5% to 15.0% higher than in 2002–2011. Under the scenario RCP 8.5, number of claims is expected to be from 1.3% to 10.0% higher than in the baseline period.
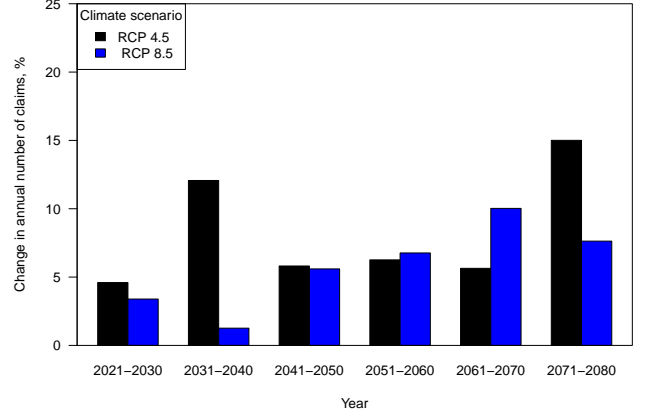
(a) Number of claims
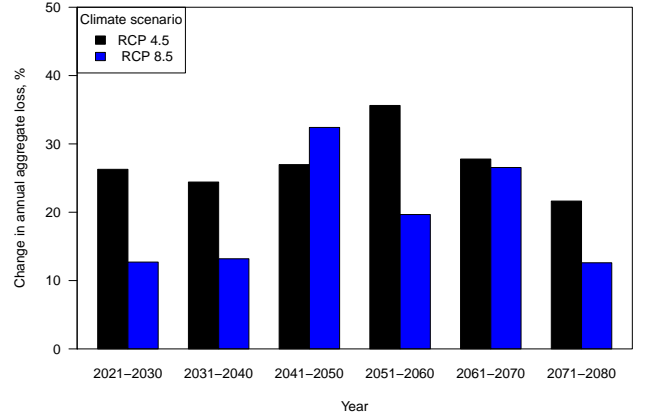
(b) Aggregate loss

Figure 3: Projected percentage changes in home insurance claims and losses in City A, relative to the control period of 2002–2011.

The magnitude of projected increase in annual aggregate loss for the same city is substantially higher, which is from 21.6% to 35.6% for the scenario RCP 4.5, and from 12.7% to 32.4% for the scenario RCP 8.5 (Figure 3(b)). Such mismatch in the relative change of the number of claims and losses (i.e., total losses grow faster than the number of claims) might be attributed, for example, to an increased future number of high impact

atmospheric events such as major floods, and/or major weather-related breakdowns of outdated civil infrastructure. In general, we can conclude that the average climate-induced loss is expected to rise regardless of the Representative Concentration Pathways (RCPs) scenario.

4.2 Goodness of Fit. In this section, we validate the employed GA-SVR models by comparing their goodness of fit with the following highly competitive approaches:

- SVR without automatic selection of the optimal hyper-parameters,

- Artificial Neural Network (ANN).

A neural network is a two-stage regression model, typically represented by a network diagram. The central idea is to extract linear combinations of the inputs as derived features, and then model the target as a nonlinear function of these features. Let X_1, X_2, X_3, and X_4 be the four input variables (like in the loss model (4.7)), Z_1 and Z_2 be the two hidden units, and Y be the output. A network diagram for this model is shown in Figure 4, where

$$Z_j = f\left(\sum_{i=1}^{4}\alpha_{ij}X_i + \alpha_{oj}\right), \qquad f(v) = \frac{1}{1+e^{-v}},$$

$$Y = g\left(\sum_{j=1}^{2}\beta_j Z_j + \beta_o\right), \qquad g(v) = v.$$

The activation functions $f(\cdot)$ and $g(\cdot)$ used in this study are sigmoid and identity activation functions, respectively. The unknown parameters (weights) α_{ij}, α_{oj}, β_j, and β_o are estimated by back-propagation method, with gradient descent updated iteratively [2], [11].

Table 3 shows that root mean squared error (RMSE) of GA-SVR approach for modeling the number of claims, N_t, is much smaller than of ANN or SVR. A substantial advantage of the GA-SVR is that the procedure better captures the variability of observed data, particularly, the sudden high spikes in the number of claims, which are the paramount concern for insurance companies (Figure 5(a)).

SVR-GA delivers at least 70% lower root mean squared error (RMSE) for the number of claims than the competing approaches, ANN and SVR (see Table 3). Furthermore, GA-SVR model also yields the most accurate fit to the observed aggregate losses, L_t, i.e., RMSE of GA-SVR is up to 15 times lower than of RMSE of ANN and SVR. Finally, GA-SVR also captures variability of the observed total losses (see Figure 5(b)).
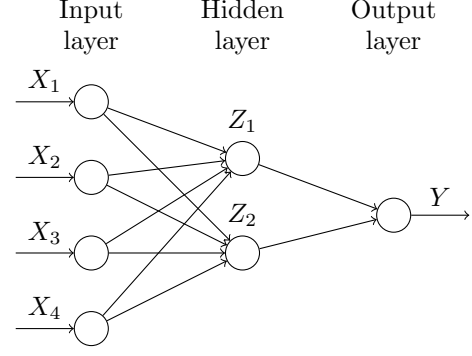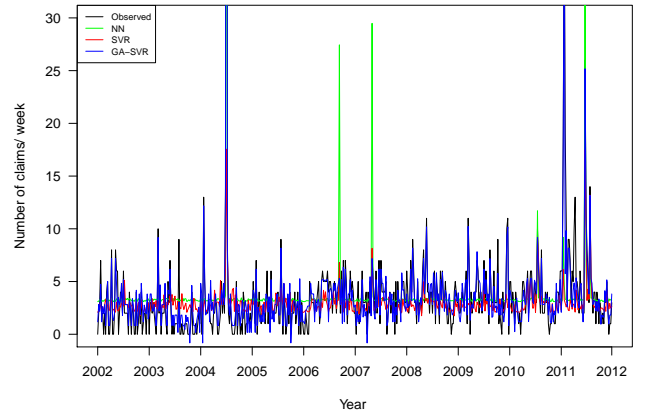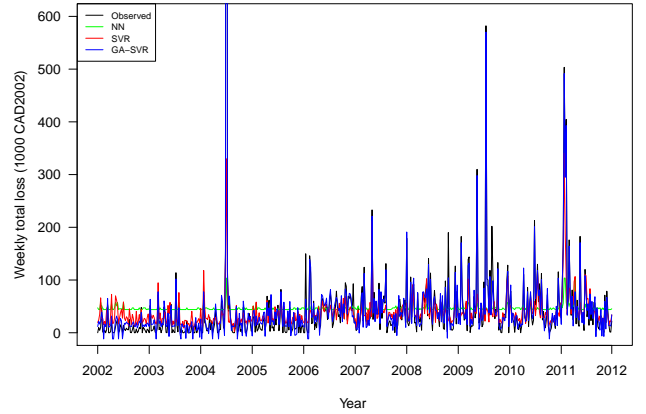

Figure 4: Neural network for four inputs and one hidden layer with two units.

(a) Number of claims

(b) Total loss

Figure 5: Observed and fitted with three different models weekly time series of home insurance claims and losses in City A.

5 Conclusion.

The purpose of this study is to model and forecast dynamics of weather-induced home insurance claims and

Table 3: Training RMSE of the models for the number of claims (N_t) and aggregate loss (L_t)

Model	N_t	L_t
ANN	5.25	212858.4
SVR	7.32	200660.1
GA-SVR	1.37	13955.6

total losses in respect to changes in precipitation. We employ support vector regression with genetic algorithm (GA-SVR)—a machine learning procedure that allows for flexible data-driven characterization of claim dynamics as a function of atmospheric information. We illustrate our approach in application to insurance claims in a middle-sized city in the Canadian Prairies.

Our model is based on a two-stage approach, that is, we first model a number of claims as a function of precipitation and then assess dynamics of total losses as a function of claim frequency and precipitation. We evaluate sensitivity of projected claims in respect to different greenhouse effect scenarios (RCP 4.5 and RCP 8.5). The projected number and magnitude of claims are compared to the baseline (control) period of 2002–2011. Our results indicate that in the future we can expect that both number of claims and insurance losses will increase in City A, but the rate of increase depends on the climate change scenarios.

It is important to emphasize that the obtained projections of future number of claims and losses account only for changes in future precipitation rather than other meteorological and non-meteorological factors, e.g., seasonal component, socio-economic changes, location and value of assets, that could also influence the dynamics of insurance claims. In the future, we plan to incorporate such information into assessment of future claim dynamics.

Predictions of future insurance claims and losses embody various sources of uncertainty. First, there is the uncertainty inherent in the climate model projections. Second, there is uncertainty about the specified models for insurance claims and losses. The third source of uncertainty is from estimates of the model parameters. In the future, we plan to quantify different sources of uncertainty. We also plan to expand the spatial domain of our analysis by evaluating the dynamics of weather related home insurance in other cities.

References

[1] D. Basak, S. Pal, and D. Patranabis. Support vector regression. *Neural Information Processing Letters and Reviews*, 11, 11 2007.

[2] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[3] G. R. Brooks. A synthesis of geological hazards in canada, bulletin 548. Technical report, Geological Survey of Canada, 2001.

[4] c2es. Heavy Precipitation and Climate Change. Technical report, The Center for Climate and Energy Solutions.

[5] A. M. Caldeira, W. Gassenferth, M. A. S. Machado, and D. J. Santos. Auditing vehicles claims using neural networks. *Procedia Computer Science*, 55:62–71, 2015. 3rd International Conference on Information Technology and Quantitative Management, ITQM 2015.

[6] K.-Y. Chen and C.-H. Wang. Support vector regression with genetic algorithms in forecasting tourism demand. *Tourism Management*, 28:215–226, 02 2007.

[7] S. Chen, K. Jeong, and W. K. Hardle. Support vector regression based garch model with application to forecasting volatility of financial returns. *SSRN Electronic Journal*, 02 2008.

[8] C. S. Cheng, Q. Li, G. Li, and H. Auld. Climate change and heavy rainfall-related water damage insurance claims and losses in Ontario, Canada. *Journal of Water Resource and Protection*, 4:49–62, 2012.

[9] M. Ghorbani, G. Zargar, and H. Jazayeri-Rad. Prediction of asphaltene precipitation using support vector regression tuned with genetic algorithms. *Petroleum*, 2(3):301 – 306, 2016.

[10] D. E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, New York, 1989.

[11] T. J. Hastie, R. J. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2 edition, 2009.

[12] O. Haug, X. K. Dimakos, J. F. Vardal, M. Aldrin, and E. Meze-Hausken. Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, 1:1–20, 2011.

[13] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2003.

[14] IBC. Water-Related Damage. Technical report, Insurance Bureau of Canada, 2017.

[15] IPCC. Towards new scenarios for analysis of emissions, climate change, impacts, and response strategies. Technical report, IPCC Expert Meeting Report, 2007.

[16] IPCC. Climate Change 2014: Synthesis Report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (R. K. Pachauri & L. A. Meyer (Eds.)). Technical report, IPCC, 2014.

[17] X. Ju, M. Cheng, Y. Xia, F. Quo, and Y. Tian. Support vector regression and time series analysis for the forecasting of bayannur's total water requirement. *Procedia Computer Science*, 31:523–531, 2014. 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014.

[18] P. Lin, S. Su, and T.-T. Lee. Support vector regression performance analysis and systematic parameter selection. *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, 2:877–882 vol. 2, 2005.

[19] V. Lyubchich and Y. R. Gel. Can we weather proof our insurance? *Environmetrics*, 28(2):e2433, 2017.

[20] NCEI. Climate monitoring. Technical report, National Centers for Environmental Information, 2017.

[21] PSEPC. Canadian disaster database version 4.4. Technical report, Public Safety and Emergency Preparedness Canada, 2006.

[22] I. Scheel, E. Ferkingstad, A. Frigessi, O. Haug, M. Hinnerichsen, and E. Meze-Hausken. A Bayesian hierarchical model with spatial variable selection: the effect of weather on insurance claims. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1):85–100, 2013.

[23] L. Scrucca. Ga: A package for genetic algorithms in r. *Journal of Statistical Software*, 53:1–37, 04 2013.

[24] S. N. Sivanandam and S. N. Deepa. *Introduction to genetic algorithms.* Springer, 2008.

[25] SOA. Determining the impact of climate change on insurance risk and the global community. Technical report, American Academy of Actuaries, Society of Actuaries, 2014.

[26] M. Soliman, V. Lyubchich, Y. R. Gel, D. Naser, and S. Esterby. Evaluating the impact of climate change on dynamics of house insurance claims, 2015.

[27] F. Tay and L. Cao. Application of support vector machines in financial time series forecasting. *Omega*, 29:309–317, 02 2001.

[28] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian].* Nauka, Moscow, 1974.

[29] V. Vapnik, S. Golowich, and A. Smola. Support vector method for function approximation, regression estimation and signal processing. In M. C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1997.

[30] S. Wu and A. Akbarov. Support vector regression for warranty claim forecasting. *European Journal of Operational Research*, 213:196–204, 2011.

[31] F.-C. Yuan. Parameters optimization using genetic algorithms in support vector regression for sales volume forecasting. *Applied Mathematics-a Journal of Chinese Universities Series B*, 2012:1480–1486, 2012.