Received: 18 September 2018 | Revised: 18 February 2019 | Accepted: 19 February 2019

DOI: 10.1002/wics.1462

WIREs COMPUTATIONAL STATISTICS WILEY

ADVANCED REVIEW

Insurance risk assessment in the face of climate change: Integrating data science and statistics

Vyacheslav Lyubchich[1] | Nathaniel K. Newlands[2] | Azar Ghahari[3] | Tahir Mahdi[4] | Yulia R. Gel[5]

[1]Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, Maryland

[2]Science and Technology, Agriculture and Agri-Food Canada, Summerland Research and Development Centre, Summerland, British Columbia

[3]Department of Statistics, Boston University, Boston, Massachusetts

[4]Actuarial and Forecasting Unit, Agriculture and Agri-Food Canada, Ottawa, Ontario

[5]Department of Mathematical Sciences, University of Texas at Dallas, Richardson, Texas

Correspondence
Vyacheslav Lyubchich, Chesapeake Biological Laboratory, University of Maryland Center for Environmental Science, Solomons, MD.
Email: lyubchich@umces.edu

Funding information
National Science Foundation, Grant/Award Number: 1739823

Local extreme weather events cause more insurance losses overall than large natural disasters. The evidence is provided by long-term observations of weather and insurance records that are also a foundation for the majority of insurance products covering weather related damages. The insurers around the world are concerned, however, that the past records used to assess and price the risks underestimate the risk and incurred losses in recent years. The growing insurance risks are largely attributed to climate change that brings increasingly more alterations and permanent impact on all aspects of human life and welfare. From floods to hail to excessive wind, adverse atmospheric events are a poignant reminder of how vulnerable our society is across a broad range of threats posed by environmental extremes. Indeed, as climate change effects become more pronounced, we face a new era of risk with increasing weather related damages and losses. This in turn, coupled with challenges of massive climatic data, requires developing innovative analytic approaches that transcend traditional disciplinary boundaries of statistical, actuarial and environmental sciences. Nevertheless, the multidisciplinary nature of climate risk assessment and its impact on insurance is often overlooked and neglected. We highlight the most recent developments and interdisciplinary perspectives on diverse statistical and machine learning methodology for modeling and assessing climate risk in agricultural and home insurances, with a particular focus on noncatastrophic events.

This article is categorized under:
Applications of Computational Statistics > Computational Climate Change and Numerical Weather Forecasting
Statistical and Graphical Methods of Data Analysis > Multivariate Analysis
Data: Types and Structure > Massive Data

KEYWORDS
agricultural insurance, climate risk assessment, environmental statistics, home insurance, uncertainty quantification

1 | INTRODUCTION

Insurance is one of the most widely recognized methods for management of loss from weather-induced risks. Those could be damages to the property (e.g., private houses, commercial buildings, vehicles, or agriculture crops) caused by such weather events as high wind, drought, cold snap, heavy snowfall or rain (Curry, Weaver, & Wiebe, 2012; Erhardt, 2017; Hall, 2017). Weather events can be characterized as catastrophic (infrequent, extreme-impact events) or noncatastrophic (frequent, low-

impact events). The latter tend to bring more cumulative losses than natural disasters (see Scheel et al., 2013, and references therein), but yet receive noticeably less attention in the literature. The current review paper primarily focuses on such low-individual but high-cumulative impact weather events that still remain a gray research zone.

Measuring the impact of noncatastrophic weather is a critical step toward raising public awareness of weather hazards, developing efficient weather risk mitigation strategies, and enhancing societal resilience (Stulec, 2017; Toeglhofer, Mestel, & Prettenthaler, 2012). For instance, noncatastrophic weather risk may have a direct impact on increasing costs and decreasing the sale volume in agriculture and construction companies (Pres, 2009). The high risks also imply more expensive insurance or even a denial of insurance coverage and, as a consequence, inability to refinance mortgages or sell a house (see Lyubchich & Gel, 2017, and references therein).

An important consideration for insurance risk management is the effect of climate change, which has been shown to amplify the frequency and severity of extreme weather events. Climate change increases the risk of weather-related damages that impact virtually all sectors of the economy, from fisheries and agriculture to tourism (Smith & Katz, 2013; Smith & Matthews, 2015).

Vast amounts of data have spurred the interdisciplinary field of data science that offers scientific methods, processes, algorithms and systems to extract knowledge and insights from data. In particular, from "big data," which are typically too large to be handled by conventional data processing software and are associated with "five Vs" characteristics: *volume*, *variety*, *velocity*, *veracity*, and *value*. Data are being collected, recorded, stored, shared, and analyzed in different ways and quickly driven by more widely-available digital technologies and increasing computer power. Such advancements offer many potential benefits for both insurers and policyholders in the design of more user targeted insurance products and claim management. Major aspects of data science for insurance are: (1) better consumer targeting and product design, (2) more accurate risk assessment, underwriting, and pricing, (3) stronger engagement with consumers, and (4) better claims management (IFoA, 2017). Alongside such potential benefits for the insurance industry, public interest and consumer expectations in the era of data science are also expected to increase. Insurance may become unavailable for some, may lead to less pooling of risk (i.e., sharing risks among policyholders having similar risk characteristics) and to reducing the insurer price discrimination where premiums are based on broader factors of consumer sensitivity. Further issues may be related to ownership of the data and associated analytics, data protection, privacy intrusiveness, public transparency, and over-reliance on the data and automated algorithms, including cyber risks. These broad benefits and challenges are associated with both agricultural and residential (i.e., home) insurance.

Anticipated climate change introduces another layer of complexity in the hierarchy of the myriad interlinked aspects and consequences of data science for risk assessment. Insurance risk assessment in the face of climate change requires a fundamental understanding of statistical methodology advancements and application insights. Despite their benefits, advanced data science methods at the interface of statistics and machine learning have not penetrated the insurance industry yet, particularly, the insurance sector associated with climate risks. One of the goals of this paper is to show how and when such interdisciplinary statistics and machine learning tools have helped to quantify climate risk in insurance.

We provide an overview of statistical peer-reviewed literature concerning insurance risks due to noncatastrophic weather events, with a particular attention to the studies assessing the effects of climate change. Sections 2 and 3 of the overview consider separately the areas of agricultural and home insurance. Analysis of agricultural risks includes analysis and prediction of crop yields, which is the area of research with long history that cannot be comprehensively accommodated in a section of this paper. Hence, we limit the discussion of agricultural studies by giving our perspective of the most recent developments in the field. Section 4 describes the key statistical approaches and challenges. Concluding remarks are given in the last section.

2 | AGRICULTURAL INSURANCE

2.1 | Perspective

Noncatastrophic weather risks may have either a direct or both direct and indirect impact on a company's finances. A direct impact usually causes a decrease in volume of sales or creates additional costs. Sales volume and price may significantly change the final total revenues and total variable costs, and often price is correlated with short-term weather conditions in a nonlinear way (Pres, 2009).

Index insurance products are essentially derivatives and are fundamentally one form of risk management that is primarily used to hedge against the risk of a contingent loss. It involves equitable transfer of the risk of loss, from one entity to another, in exchange for a premium that is charged for a certain amount of insurance coverage. Agricultural insurance is an important part of ensuring long-term stability and growth of the agriculture sector, and facilitating access to credit, helping to reduce the

negative impacts of natural catastrophes, and encouraging investment in improved production technology (Porth & Tan, 2015).

Insurance rate-making is complicated in two major ways. The first way is due to information asymmetry (adverse selection) and moral hazard due to insurance fraud, when potential policyholders have proprietary knowledge about their risk exposure that is not available to the insurer, or when insurance underwriters assign potential policyholders into risk rating classes, but misclassify them due to incomplete information. Once insured, policyholders may take on greater risk, but unless an insurer can effectively monitor the policyholder's behavior to enforce policy provisions, insurers can experience losses that exceed the projections used to establish premium rates and be forced to increase premium rates for all. The second way is due to basis risk when policyholders cannot get compensation for actual losses, or are paid an indemnity with no actual loss. Basis risk occurs when the underlying risk is not perfectly correlated to the actual loss (due to confounding factors, spatial and temporal issues), whereby an index must be able to explain a very high portion of the variability in losses to reduce basis risk. While indemnity contracts are paid according to a farmer's actual losses, weather index-based insurance is paid based on some index level. This assumes an index is highly correlated to actual losses to effectively overcome information asymmetry, avoiding adverse selection, and moral hazard. It is crucial to minimize basis risk to ensure wide insurance adoption and coverage, even if premium rates are reliable and robust. Nonetheless, it is impossible to fully eliminate this risk due to inherent, irreducible uncertainty between actual losses and a measured peril when a policy includes multiple farmers within a region having different loss severity and extents of the coverage.

Historical crop yield data at the farm level is often insufficient or unavailable, hence, unreliable for estimating individual expected losses, due to lack of broad representation and potential selection bias. This is attributable to both data scarcity and credibility at the farm or field scale. Data scarcity occurs because farmers apply crop rotation that creates gaps and inconsistencies in time series of yield. Credibility issues occur because of interannual changes in both technology (includes water, fertilizer, and pesticide inputs) and cropping practices, so that historical data is no longer representative of current practices. Aggregation bias can occur if aggregated regional-scale data (at the level of county or municipality) are used in place of farm-scale data. It may also lead to potential cancellation of idiosyncratic risk between different assets and, in turn, underestimation of predicted risk relative to true risk (Gerlt, Thompson, & Miller, 2014; Porth, Tan, & Zhu, 2016).

Despite great effort to apply statistics in developing and designing robust insurance rate-making mechanisms or schemes, there are many reasons insurance markets simply fail—such as lack of legal systems to enforce contracts, insufficient market infrastructure, lower producer risk awareness, lack of insurance culture, strong covariate risk exposure, and high transaction costs that make insurance coverage prohibitively expensive, relative to its benefit (Porth & Tan, 2015). Almost all crop insurance programs involve a public-private partnership approach, supported by government subsidies from tax revenue because of these insurability challenges (Smith & Glauber, 2012). Since governmental support is used extensively in agricultural insurance, many related data sets are publicly available, whereas data sets related to other types of insurance (home, auto insurance, etc.) are usually proprietary.

In Canada, agriculture insurance programs are delivered in partnership with the provinces. Most of the provinces have created crown corporations that deliver agriculture insurance programs within the province, while in some provinces, the provincial government itself delivers the programs. Crop insurance delivery costs are fully subsidized by provincial and federal governments each paying 50% of the delivery costs, with provincial crown corporations or the province delivering crop insurance, and the federal government contributing a portion of total premiums and administrative costs through its AgriInsurance program. Premiums for crop insurance are subsidized by stakeholders differently at three funding levels: catastrophic, comprehensive, and high cost (Table 1). In the United States, crop insurance is delivered through private insurance companies, with premiums subsidized on average at 60% (federal government), and the remaining 40% paid by producers.

Agricultural insurance plans can be: single-peril, providing coverage against one peril or risk, or multiperil. The perils typically covered under an agriculture insurance plan are: excessive moisture, wind, diseases, frost, excessive rainfall, excessive heat, drought, and flooding. Multiperil insurance plans can be further divided into yield based plans, nonyield based plans, revenue insurance plans, income based insurance plans, whole-farm, area-yield index-based or area-revenue index-based, weather derivative plans, and plans covering trees and vines mortality. The indemnity rate of an agricultural product (crop)

TABLE 1 Percent shares of the stakeholders of the crop insurance programs in Canada

	Stakeholder level		
Funding level	**Federal**	**Provincial**	**Producer**
Catastrophic	60	40	0
Comprehensive	36	24	40
High cost	20	13	67

moves up and down every year depending upon many risk factors such as weather, plant diseases, underwriting inefficiencies, technological changes, and management practices. Underwriting process can be improved by accounting for the following:

- The insurance coverage for a producer can be a function of producer's own yields as well as provincial or regional yields. Regional yields may not be a good indicator of a particular producer's yields and may introduce bias in calculation of coverage for some producers.
- Farmers rotate crops, thus, the seeded acres of a crop and productive capability of all acres may not be same every year. When the coverage for a producer is calculated based on the historical yields that are not on the same acres, it creates bias.
- Due to crop rotation, farmers may not seed a particular crop on their farm and may not have actual yield for that crop in a particular year. This opens the door for the use of regional yields for the years where the producer does not have his own yields and regional yields may not be indicative of the productive capability of the producer.
- Typically acres in a farm are not homogeneous. The productive capability of acres in a farm varies depending upon the level (high and low land) and type of soil among other things. If coverage can be calculated for each acre or homogeneous blocks of acres, the coverage calculation would be more accurate.
- Coverage is currently calculated based on yields. With improved technology, weather-related variables can also be used to calculate coverage. Historical insurance records show some losses even in bumper years when the weather was suitable. Some of those losses are due to inefficiencies in coverage calculations. With the use of machine learning, improved technology and software, some of the underwriting inefficiencies can be controlled or eliminated.

The above inefficiencies are drawn from Canadian actuarial situation, but can occur also in other insurance programs around the world.

2.2 | Data credibility issues in premium rates development

Premium rates are based on the indemnity rates observed in the past years. In Canada, the premium rates are calculated at the regional (risk areas) or, more often, at the provincial level. The risk areas used to develop premium rates are quite large and may not reflect the risk profile of smaller areas correctly (AFSC, 2018).

At the same time, the loss data for smaller areas is insufficient (not credible) to calculate premium rates. It creates inequity between producers as some producers pay higher premiums than they would if data for smaller areas was credible. Climate and soil quality data can be used to supplement the available data so the lack of credibility can be resolved.

2.2.1 | Additional comments

Noncatastrophic weather events have caused more insurance losses in agriculture than catastrophic weather events. There have not been province-wide catastrophic events in Canada in the last 15 years. However, there were small-scale catastrophic events. Examples include: Heat waves in Ontario, Saskatchewan, and Alberta caused large losses in some small areas but could not be considered catastrophic at the provincial level. In March 2012, air temperatures in parts of Ontario were higher than normal and fruit trees bloomed earlier than usual, but the flowers were lost due to a frost later. As a result, the fruit trees did not have fruit that year. This event was catastrophic for fruit producers but not significant at the provincial level (Environment Canada, 2017).

Similarly, the frequency and length of heat waves is increasing in the USA, strengthened by long-term drought conditions. In summer 2011, heat waves affected Southern Plains rendering majority of pastures in Texas and Oklahoma as very poor for most of the growing season. Drought and heat waves in North Dakota, South Dakota, and Montana caused damage to crops and lack of feed for cattle forcing the ranchers to sell off livestock. In March 2017, many crops have been already blooming in southeastern states of the USA due to unusually warm weather when a sudden 2-day freeze caused severe damages to peaches, blueberries, strawberries, and apples among other crops. The most impacted states were Georgia and South Carolina, and the overall losses over nine affected states were estimated at about one billion U.S. dollars (NOAA, 2018).

Other recent examples include heat waves in Europe and Asia in summer 2018, which were rather extreme, hitting long-term temperature records and causing severe damage to crops and increased prices of crops and of grain milling (Patel, 2018).

2.3 | Weather derivatives

Weather derivatives are financial instruments that can be used by organizations or individuals in a risk management strategy to calculate indemnity payments using weather variables such as amount of rainfall, number of corn heat units, temperature, humidity, or snowfall (Alexandridis & Zapranis, 2013). Generally, weather derivative plans include hay, forage, and pasture

insurance plans, but can also be used for traditional yield-based multiperil plans. Weather variables are a valuable tool to estimate payments. Furthermore, the use of weather variables reduces underwriting costs that might be significant.

Weather derivative plans are based on weather data collected at weather stations. The number of weather stations is limited and sometimes the weather stations are far from some of the insured fields and may not accurately characterize the situation on the farm. This creates the basis risk discussed earlier.

Currently, most of the weather derivative plans use one weather variable that is linked to the production and is used for indemnity calculations (e.g., AFSC, 2018). Using simply one weather variable as an indicator of production may not be the best way of estimating production. With the help of advanced multivariate statistical and machine learning tools accompanied by high-resolution climate data, application software can be developed that can be used by the underwriters to estimate yields remotely and with higher accuracy.

2.4 | Statistical advancement addressing climate change

In addition to weather risks affecting spatial and temporal correlation of insurance indices and their covariates, climatic change also can increase the volatility of weather variables, generating nonstationary loss distributions, which are challenging to estimate reliably, adding uncertainty to actuarial rate-making (Odening & Shen, 2014). Shiraishi (2016) provides an in-depth mathematical review of alternative insurance risk models and the ruin probability (i.e., probability that an insurer's surplus is below zero) along with its representation and estimation using the Gerber–Shiu function in the context of dividend, reinsurance, and taxation applications. We provide a broader, multidisciplinary review of advancements in the application of statistical tools and actuarial methods that support weather/climate index-based risk assessment and improved design of insurance products. Advancements in monitoring and surveillance infrastructure (such as weather stations, sensor networks, and remote sensing technologies) and increases in computing power have enabled the development of new statistical and mathematical models.

Remotely sensed (RS) data, such as satellite-based rainfall estimates, help to improve and validate the indices, in addition to being used operationally to track insured seasons and assess basis risk (Black, Greatrex, Young, & Maidment, 2016). While RS has fewer issues with data sparsity, it is still only a proxy for actual areal average rainfall and may not be representative of local conditions experienced by an individual farm. In the review of the potential and uptake of RS in insurance, de Leeuw et al. (2014) highlight the need for greater cooperation and multidisciplinary understanding to develop RS indices tailored to the value and profitability of the insurance industry. The review also indicates a need to develop indices that are best correlated with what is insured, that can be delivered at sufficiently low cost, and that target emerging markets other than classical claim-based insurance (de Leeuw et al., 2014). Leblois and Quirion (2013) provide a general overview of the methods used and difficulties faced by weather-based insurance indices (rainfall, water stress, and drought) and RS vegetation indices, namely leaf area index (LAI) and normalized difference vegetation index (NDVI). Leblois and Quirion (2013) point out that simple detrending methods based on past data are typically used, but such methods do not correctly account for complex nonstationarities in space and time linked with temperature or rainfall.

Cluster analysis of rainfall to determine the trigger for signaling a drought-insurance payout in Ghana was applied by Choudhury, Jones, Okine, and Choudhury (2016). They demonstrate the feasibility of a rainfall-based index insurance product that minimizes moral hazard and adverse selection risk, while enabling a more rapid and structured payout process. Nonetheless, the over-reliance on historical climate observations to guide the design of insurance products by incorporating climate variability and climate change trends may lead to premiums that mislead both policyholders and insurers. Incorporating bias-corrected climate model output and climate reanalysis data into the pricing decisions may further exaggerate premiums and viability of weather index-based insurance, as demonstrated using Bayesian belief networks (Daron & Stainforth, 2014). Recent findings on the effect of large-scale patterns of atmospheric pressure and circulation anomalies (also known as climate teleconnections, such as El Niño southern oscillation, ENSO) on US government-set premium rate pricing for cotton, using a moment-based maximum entropy modeling approach, show, for example, that private insurance companies could reduce their paid indemnities by 10–15% on average (Tack & Ubilava, 2015). Furthermore, Dalhaus and Finger (2016) find that there is no difference between using gridded climate reanalysis and weather station precipitation data, rejecting the hypothesis that precipitation grid data reduce spatial basis risk, but instead report that observational data on crop development (phenology) can significantly increase expected utility (Conradt, Finger, & Spörri, 2015; Dalhaus & Finger, 2016). Also, Woodward (2016) has recently investigated the integration of high-resolution soil data from the state of Illinois, USA, into crop insurance policy design, reporting that the degree to which soils vary within a county is highly significant, leading to rating errors of 200% or greater. Crane-Droesch (2018) uses crop data from Illinois and other states in the US Midwest and opts for using gridded meteorological data. Crane-Droesch (2018) suggests to employ features obtained from the last layer of a deep neural network in a parametric statistical model for panel data, thus, merging the two techniques into a semiparametric neural network. Newlands, Ghahari, Gel, Lyubchich, and Mahdi (2019) use a variety of machine learning methods, including random

forest and two versions of deep neural networks, to demonstrate the aptness of machine learning methods and comparative utility of station-based and gridded climate reanalysis data in forecasting crop yields over a large territory.

Recent studies have investigated the utility of copulas in modeling dependence, better accounting for extremes and collective, agent-based dynamics. Copulas are multivariate probability distributions for the dependence between random variables for which the marginal probability distribution of each variable is uniform. They are a statistical measure or function for examining the dependence between many variables, beyond strict assumptions of linearity and normality. Unlike the Pearson's linear correlation, copula is invariant under monotone increasing transformations of relevant random variables. Copulas have been used to quantify systemic weather and yield risks, and to develop local test procedures for detection of weather risk changes, by augmenting observational data using expert knowledge in a Bayesian estimation framework (Ahmed & Serra, 2015; Odening & Shen, 2014). Odening and Shen (2014) studied revenue insurance providing joint price and yield coverage, to guarantee a minimum income for apple and orange growers in Spain. The findings indicate the potential of copula-based mixtures to perform better than individual copulas, and that a mixture between Archimedean copulas is capable of improving insurance pricing. The standard Gaussian copula model is shown to significantly underprice risk, due to disregarding tail dependence, which is a critical factor when risks are state-dependent (Goodwin, 2015).

Conradt, Finger, and Bokusheva (2015) developed a flexible insurance design based on quantile regression (QR) to condition yield index dependency, extending the classical mean-conditioned view to the extremes. The approach is validated for wheat using three statistical risk measures, namely expected utility, expected shortfall, and spectral risk measure, showing that QR represents lower tails of the distribution better than ordinary least squares (Conradt, Finger, & Bokusheva, 2015).

With climate change, increasing weather extremes induce more severe crop heat stress that reduces crop yield. In a recent assessment of present (1975–1990) conditions and future (2030–2050) regional climate model scenarios in seven sites for growing wheat in the Mediterranean basin, expected insurance payouts were found to increase, when crop yield is modeled using a negatively-skewed Weibull probability density function. Payouts increased on average between 11 and 25%, relative to unstressed conditions further supporting the need for flexible risk management strategies (Moriondo et al., 2016). Furthermore, applying a population dynamical approach involving risk sharing among individual agents among collectives in a *N*-person threshold game shows that collective index insurance offers clear advantages over individual index insurance as a viable, resilient insurance strategy (Pacheco, Santos, & Levin, 2016).

Two recent reviews of machine learning techniques in agriculture (Kamilaris & Prenafeta-Boldu, 2018; Liakos, Busato, Moshou, Pearson, & Bochtis, 2018) evidence that the data science methods are underutilized in agricultural risk assessment, such as yield prediction. Among the studies mentioned in the reviews, Kuwata and Shibasaki (2015, 2016) bring together gridded weather RS data and deep neural networks to estimate corn yields in a current year. Pantazi, Moshou, Alexandridis, Whetton, and Mouazen (2016) use supervised Kohonen networks, counter-propagation artificial networks, and XY-fusion networks to forecast winter wheat yield in 2013 on a small scale—22 ha in Bedfordshire, U.K., using NDVI and a number of soil parameters that are typically unavailable for insurance companies (soil moisture content, pH, and nutrient concentrations). Kung, Kuo, Chen, and Tsai (2016) propose to combine several neural networks in an ensemble to improve predictions of tomato yields based on a number of meteorological factors. Overall, forecasting results that are obtained with machine learning for larger geographical areas and could be used by insurance companies are yet scarce (see Table 2 combining information on these and the earlier mentioned studies).

2.5 | Knowledge gaps and challenges

In summary, the review identifies the following set of questions, needs, and knowledge gaps:

- While weather factors are more relevant to crop yield uncertainties than soil variations across larger area (district, province), there is still a need to assess the performance of soil-landscape suitability-based and soil moisture-based index insurance for real farms in a comparable moderate region on a whole-farm scale (Doms, 2017).
- Crop phenological observations need to be further explored to improve insurance index products and offer more flexibility to reference different critical development and growth stages (phenology phase accumulation indices). Such flexibility can reduce farmers' downside risk exposure.
- Most index products are based on a single variable such as rainfall, county yield, and satellite-based NDVI values, but complex weather-based products need to be explored and validated to insure against multiple events.
- There shall be developed curated archives of gridded climate data at temporal and spatial resolutions applicable to insurance practice. Such data availability could significantly increase the attractiveness of weather index-based products, without providing premium subsidies (Dalhaus & Finger, 2016).

TABLE 2 Studies on noncatastrophic weather risks for agricultural insurance

Study	Insurance data	Major methods	Important correlates	Risk forecast
Ahmed and Serra (2015)	Annual Spanish average prices and yields for apple and orange; 1954–2010	Statistical copulas	Commodity price, revenue, yield	Estimation of expected losses and fair premium rates at 75% and 80% coverage (no prediction)
Choudhury et al. (2016)	Yield data across the northern region of Ghana (1994–2007)	Multiple regression, autoregressive error model, model-based cluster analysis	Climate change cycle, rainfall	Estimation of rainfall trigger and stop-loss estimates (R^2's 88–97%) for four districts of Ghana based on training data from 14-year historical period (no prediction)
Conradt, Finger, and Bokusheva (2015)	Yield data from 47 individual farms in five counties of Northern Kazakhstan; data source: regional statistical offices of Kazakhstan (1980–2010)	Quantile regression	Cropping area, rainfall	Estimation of expected utility, expected shortfall, spectral risk measures for 47 single-farm wheat yield data from Northern Kazakhstan based on 31 years long time series (no prediction)
Conradt, Finger, and Spörri (2015)	Farm-level panel data on wheat production of 47 single farms of 5 different counties in Kazakhstan (1980–2009)	Quantile regression accounting for phenological plant growth phases	Growing degree days (GDD), start and end of the insured period, average sowing date, average time lengths of different wheat phases, daily temperature and precipitation values from five weather stations in each county	Evaluation of yield-weather indices (no prediction)
Crane-Droesch (2018)	County-level corn yields in nine states in the USA in 1979–2016	Deep neural network and new augmented model of neural network and parametric statistical model for panel data (semiparametric neural network, SNN)	Time, precipitation, GDD, air temperature, relative humidity, wind speed, shortwave radiation, and functions of those	Predictions for 2040–2069 and 2070–2099 based on 13 climate models and two climate scenarios (RCP 4.5 and 8.5) indicate decline of corn yields, based on all combinations of prediction models and climate projections
Dalhaus and Finger (2016)	Farm-panel data (winter wheat) for 29 farms in Germany (1996–2010)	Quantile regression	Gridded (1 km) rainfall, weather station data near 29 farms (0.45–18 km), timing of wheat phenological phases	Estimation of insurance parameters (strike level, optimal accumulation period) for gridded rainfall and crop phenology (no prediction)
Daron and Stainforth (2014)	Insurance design with climate data, Kolhapur district, Western India (1990–2008) and climate scenario data	Bayesian belief network	Gridded rainfall (1961–2004), UK Hadley Centre regional model (HadRM3) A1B scenario, 1960–2099	Prediction of premium, maximum indemnity (%) for past and future climate periods (future premium increases of 29–31% of maximum indemnity in 2030s
Gerlt et al. (2014)	County and farm-level data by National Agricultural Statistics Service (NASS), for example, 1,513 farms in Illinois, USA	Scenario-based simulation of farm yields with resampling, nonparametric kernel density estimator	Crop yield	Estimation of crop yield and average farm based premium under different coverage level (no prediction)
Goodwin (2015)	County-level yield and price data for corn and soybean within four counties, Illinois, USA (1960–2012)	Copula modeling, marginal distribution estimation	Crop yield, price	Estimated revenue insurance loss probabilities and claim rates
Kung et al. (2016)	Tomato harvest, planted and harvested area obtained from the Agriculture and Food Agency of Council of Agriculture in Taiwan, 1997–2014	Multiple regression, deep neural networks, network ensemble	Relative humidity, precipitation, planting area, air temperature, cost of production, and market trading price	Prediction error of 1.14–7.15% in a series of experiments, using ensemble network approach
Kuwata and Shibasaki (2015, 2016)	Annual corn yield at the county level in the USA in 2008–2013	Support vector machine, deep neural network	Surface minimum and maximum temperature, precipitation, humidity, shortwave radiation, snow water equivalent	Estimated crop yield index (in standard deviations) for the current year
Moriondo et al. (2016)	Crop yield (simulated) for seven sites in Mediterranean based on historical (1975–1990) and future (2030–2050) climate	Crop growth simulation (Sirius), Weibull distribution model estimation	Crop yield (durum wheat), temperature, precipitation	Expected insurance payouts increase under heat stress compared to unstressed conditions for historical (+11%), and in the future (+25%)
Newlands et al. (2019)	Municipality-level yields of 75 crop types during	Deep belief and deep neural networks, random forest, and gradient boosting	A matrix of more than 400 weather indices derived from daily observational data from	Cross-validation "one-year-out" forecasts of yields simultaneous for all municipalities

TABLE 2 (Continued)

Study	Insurance data	Major methods	Important correlates	Risk forecast
	1996–2011 in Manitoba, Canada		weather stations or from gridded weather reanalysis data	
Pacheco et al. (2016)	None	Population dynamics modeling (*N*-person threshold game)	Individual/collective coverage costs, trigger and loss probability, payout, weather index, initial individual wealth, risk tolerance, insurance profitability threshold	Theoretical prediction of utility, fitness and payoffs for risk-averse and risk-prone behavioral strategies
Porth et al. (2016)	Farm- and county-level corn yields for Canada and USA	Relational, multiscale model (farm, county)	County size, county-level yield and yield variance, average farm size, growing season monthly average temperature, cumulative precipitation	Prediction of farm-level yield distribution in country in the absence of farm-level yields with mean prediction error of 28.85% (standard deviation of 70.97%) showing aggregation bias in county-level yield data to approximate farm-level premium rates leads to underestimation of up to 40%
Tack and Ubilava (2015)	Cotton yields for 224 counties in the USA for 38 years, by National Agricultural Statistics Service (NASS), 1968–2005	Moment-based maximum entropy (MBME) modeling	Sea surface temperature (SST), El Niño Southern Oscillation (ENSO) anomaly (NINO3.4 index), rainfall	Estimation of impact of ENSO on premium rates under different coverage levels (no prediction)
Woodward (2016)	Farm-level corn yields for five counties from the Illinois Farm Business Farm Management database (FBFM), 1972–2008	Conditional Weibull distribution model estimation	Crop acreage, soil productivity rating, summer-averaged Palmer drought severity index (PDSI) (1895–2009 or 1980–2009)	Prediction of expected loss cost ratios conditional on extreme weather events (1988 and 2008) and under different technology levels

- There is a need to further explore the use of crop yield and seasonal weather forecasts in weather index-based insurance. Coupling crop yield and weather forecast indices provides probabilistic information on the next growing season and are anticipated to continue to increase in accuracy.
- Further investigation of agent-based approaches to better understand collective behavior and dynamics with respect to socio-economic benefits and challenges of insurance are needed.
- The wider availability and integration of private crop insurance data, relevant governmental databases, including RS-based crop mapping and yield indices to enable an evaluation of the performance of deep learning (i.e., the area of machine learning in which artificial neural networks adapt and learn from vast amounts of data) at the regional and field level.
- There is a demand for data science models that utilize available crowdsourced data from farmers' smartphones, and models that can utilize optical and/or synthetic aperture radar (SAR) satellite-based remote sensing data such as soil moisture and vegetation indices (LAI and NDVI) from Sentinel 1/2/3 and RADARSAT2/RADARSAT Constellation Mission (RCM) satellites.

3 | HOME INSURANCE

3.1 | Recognizing the importance and directions

A single home insurance policy often covers multiple perils or causes of loss (such as fire, water damage, and theft) except specifically excluded (e.g., earthquake damage). The research on the topic of assessing and forecasting home insurance risks from noncatastrophic weather events had to start with studies that would show the need and value of such analysis. The study of Grace, Klein, and Kleindorfer (2004) from that period takes a step in decomposing the demand for homeowners insurance into catastrophe and noncatastrophe coverage. Grace et al. (2004) use two-stage least squares to fit econometric models for the insurance demand with multiple explanatory variables, some of which are endogenous. The data come on a postal code (zip-code) level from two states in the USA, Florida and New York. The results of the analysis are consistent across the states and show that price elasticity of demand for catastrophic coverage is higher than for noncatastrophic events. This study, however, does not distinguish between noncatastrophic weather events and other noncatastrophic perils.

While the next study (Pres, 2009) is not on home insurance, we mention it here because it suggests working specifically with noncatastrophic weather risks. Pres (2009) states that noncatastrophic events cause losses only for weather-sensitive companies (implying only their direct financial results, rather than general property damages). Pres (2009) advocates for risk modeling for improved quantification of weather-related risks. For this purpose, a multivariate linear regression is suggested,

where response is a financial outcome and regressors include weather indices, polynomial time trend, and dummy seasonal variables. To overcome possible violation of ordinary least squares assumptions due to autocorrelation or heteroscedasticity of residuals, the Newey–West estimator is used to estimate standard errors of the regression parameters.

The analysis by Frees, Meyers, and Cummings (2012) offers substantially more details on different perils and further classifies noncatastrophic weather risks into separate groups of water, wind, hail damage, etc. Homeowners summary statistics (based on sample data from several major insurance companies, covering most of the United States) show that wind and water damage are the most frequent perils, whereas the highest median claims are due to hail damage (see Frees et al., 2012, Table 1).

These three studies (Frees et al., 2012; Grace et al., 2004; Pres, 2009) lay the basis for further research on assessing noncatastrophic weather risks. For example, as we show in the next section, most of the later developed analyses deal primarily with water as the major peril. In addition, a number of publications (e.g., Donat, Leckebusch, Wild, & Ulbrich, 2010, 2011; Held et al., 2013; Klawa & Ulbrich, 2003) focus on winter storm damage, including hail; however, the primary attention of these studies is rather on the very extreme events and improved projections (and their ensembles) of climatic variables, and not on statistical modeling of claims and losses. For instance, Klawa and Ulbrich (2003) analyze a series of named storms and maximum wind speeds of above 98th percentile during the reference period. Note that insurance losses for such extreme events are not a sole responsibility of insurance companies, but also taken care of by the higher level of insurance, reinsurers, such as MunichRe (Donat et al., 2011) and as such fall outside the primary scope of this overview.

3.2 | Targeting the risks

There have been several interdisciplinary research teams addressing the topic of noncatastrophic weather-related risks in home insurance, and probably the earliest results belong to the group based in the Norwegian Computing Center. Haug, Dimakos, Vardal, Aldrin, and Meze-Hausken (2011) started by modeling water damage to private buildings in Norway, using generalized linear models (GLMs) which relate numbers of claims and, separately, claim severities to weather conditions. The explanatory variables included municipality-level daily average air temperature, precipitation amounts at the current and previous day, average precipitation during the last 5 days, snow water equivalent, drainage runoff, linear time trend, and two pairs of Fourier series for seasonal components with periods of 1 year and half year. While the employed GLMs bore some simplifying assumptions, the study created an analysis framework for linking the high-quality insurance data with weather observations and for forecasting the risks and associated uncertainties using output of a global climate model (GCM) in the GLMs. Haug et al. (2011) also raise the awareness and suggest their solutions to the problems in spatial downscaling of weather variables and calibrating climate model output. Overall, this study can be used as a baseline for future research in noncatastrophic weather-induced risk insurance, providing important insights on limitations of the currently available modeling approaches, uncertainty quantification, and data availability.

The more recent studies from the Norwegian Computing Center replace the GLMs with Bayesian Poisson hurdle model, modify the list of predictors (omit linear trend and Fourier series, shorten the window for aggregated precipitation to 3 days), and eliminate long-term forecasts based on climate models. In particular, Scheel et al. (2013) provide one-week ahead forecasts (the period of 1 week is selected based on trustworthiness of the weather data used in such forecasts); Scheel and Hinnerichsen (2012) build scenario forecasts by increasing the 2001 meteorological and hydrological covariates by 5, 18, or 30%.

Cheng, Li, Li, and Auld (2012) offer a similar framework of studying the relationships between number of insurance claims and total incurred losses in the observed period and making long-term predictions based on GCM projections. Their monthly insurance data reflect personal and commercial property damages caused by rainfall in four cities in the province of Ontario, Canada. The rainfall data come from historical records and from five downscaled outputs of GCMs; historical runs of the GCMs were used to correct the biases in the GCM projections. Cheng et al. (2012) put most of their effort in creating hourly, daily, and monthly rainfall indices that account for intensity of the rainfall and potentially have a better explanatory power for the corresponding insurance outcomes. Cheng et al. (2012) notice a threshold effect in the relationships between their monthly rainfall index and insurance variables. To forecast a number of claims (or total losses) over some period, the threshold is used to compute proportions of the period with projected rainfall index below and above the threshold, then each proportion is multiplied by average observed number of claims (or total losses) in the corresponding category and added together (Cheng et al., 2012, equation (3)). With this simple approach, months within each category (below or above the threshold) are assigned the same average value (average number of claims or average losses) from insurance records. In this way the analyst has to assume that the scale of projected rainfall is not much different from the observed one, otherwise, some adjustment to the insurance quantities is desired.

The publications by a Dutch research group (Spekkers, Clemens, & ten Veldhuis, 2015; Spekkers, Kok, Clemens, & ten Veldhuis, 2013, 2014) explore rainfall-related damages to private property and content in the Netherlands, at the district level. The idea is to model claim frequency and claim sizes using four sets of predictors (rainfall-related, socio-economic, building-

related, and topographic) with GLMs and decision trees (the trees were obtained only for the frequencies, no acceptable tree could be obtained for average claim size; Spekkers et al., 2014). Similarly to Cheng et al. (2012), Spekkers et al. (2013) choose thresholds for precipitation intensity, but make no forecasts of future insurance risks.

Furthermore, Frees et al. (2012) demonstrate predictive modeling of home insurance pure premiums and frequency and severity of claims from multiple perils. Frees et al. (2012) argue that classification of perils is ambiguous and perils themselves are not independent. The suite of models they are proposing provide a flexible framework for considering multiple risks independently or as a portfolio by incorporating instrumental variables in corresponding GLMs. With 404,664 records of more than 100 variables sampled over the USA, nine perils are considered with two sets of explanatory variables: (i) policy-related variables, such as policy amount of coverage, deductibles, and age of the building, and (ii) standard insurance industry variables for weather and elevation, vicinity, commercial and geographic features, trends and ratings (Frees et al., 2012). The high number of variables used in the models makes it impossible to obtain long-term forecasts, such as based on future climate scenarios. The authors stress that the complex approach involving modeling the associations between perils might not be the best overall, since simpler approaches can be more interpretable and permitting to focus on specific perils and variable selection.

Finally, Lyubchich and Gel (2017) use publicly available daily information on number of flood-related home insurance claims in two counties in Norway, Ostfold and Sor-Trondelag, as well as observed daily precipitation and its downscaled GCM projections. They direct their efforts on improving the attribution analysis by discovering thresholds in the response of number of claims to rising precipitation (using alternating conditional expectations and machine learning approach of decision trees) and by incorporating those thresholds into generalized autoregressive moving average models (GARMA). Soliman, Lyubchich, Gel, Naser, and Esterby (2015) and Lyubchich and Gel (2017) present two new nonparametric methods to compare tails in distributions of observed and GCM-projected weather variables, using data for several Canadian cities and Norwegian counties. The methods represent an approach to numerical integration, where the integration points are equidistant on a probability scale for the quantiles (Soliman et al., 2015) or on the original scale of the variables (Lyubchich & Gel, 2017). The results of those tail comparisons can be used directly in regression. Lyubchich, Kilbourne, and Gel (2017) extend these methods to joint frequency-severity predictive modeling of insurance claims related to water damage in four Canadian cities, and run forecasts for two scenarios of greenhouse gas emissions. Overall, this research group actively combines machine learning approaches (alternating conditional expectations, regression trees) with parametric and nonparametric statistical methods (GARMA models, tail comparison and resampling), but does not select machine learning approaches as primary tools for analysis.

Table 3 combines information from the mentioned publications and shows that there are just a handful of methods that have been implemented in this line of research so far. In particular, the potential of many machine learning methods remains untapped, as well as of advanced nonparametric statistical techniques. Functions of the rainfall appear to be the most important correlates (predictors) for home insurance risks; however, it is still hard to predict future risks reliably. The risk forecasts vary dramatically based on statistical and climate models, as well as based on climate scenarios forcing those climate models. Across the studies, we see percentage increases of about 0–6,800% in the future claims and losses, relative to some baseline period in the recent past. The point on which all forecasts agree is that there will be a general increase of home insurance risks that needs attention of the industry and the public.

4 | KEY METHODOLOGICAL APPROACHES AND CHALLENGES

Most considered models for insurance risks Y_i (such as crop yield, insurance losses, or number of claims) are based on the general regression framework:

$$\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{Y}$ is an $n \times 1$ column vector comprising observations of the variable representing insurance risk; $\boldsymbol{\mu}$ is an $n \times 1$ column vector of expected values $\mathrm{E}(Y_i) \equiv \mu_i$; $\boldsymbol{\epsilon}$ is an $n \times 1$ column vector of zero-mean random deviations from the expected values, $i = 1, \ldots, n$, and n is the sample size.

In case of a multiple linear regression, the mean response takes the form

$$\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where $\mathbf{X}$ is an $n \times (d + 1)$ matrix with one column of 1's for fitting an intercept in the model and the remaining d columns for d correlates (i.e., explanatory variables, such as precipitation amounts, number of past claims, or day of year) associated with the risk variable; $\boldsymbol{\beta}$ is a $(d + 1) \times 1$ column vector of regression coefficients.

The estimation of regression model (2) and further inference are based on a *number of assumptions* about the validity of the form of the model (i.e., linearity of relationships between Y and each X variable), linear independence of the

TABLE 3 Studies on noncatastrophic weather risks for home insurance

Study	Insurance data	Major methods	Important correlates	Risk forecast
Cheng et al. (2012)	Monthly claims and losses caused by water damage, rainfall-related sewer backup, flood and business interruption; four cities in Canada; 1992–2002	Rainfall index construction	Monthly rainfall index	Five GCM-based predictions for 2016–2035, 2046–2065, and 2081–2100. Percentage increase varies from about 3 to 50%
Frees et al. (2012)	Sample of more than 400,000 U.S. policyholder year records with information (whether there were any claims and the amount associated) on 9 perils	Structural (simultaneous) regression model for each peril $i = 1, \ldots, 9$ with instrumental variables for 8 other perils $j = 1, \ldots, 9$; $i \neq j$	"Basic" set of policy-related variables: coverage amount, building adjustment, construction age, policy form and deductibles, and base cost loss costs (proxy for territory)	—
Haug et al. (2011)	Monthly data on number of private buildings insurance policies and daily data on claim frequency and severity due to local precipitation and melting snow; all mainland municipalities of Norway; 1997–2006	Separate GLMs for probability of claim and average claim size	Daily precipitation	Predictions for 2071–2100 based on two climate change scenarios from HadAM3H climate model. Percentage increase for three selected counties is within 0–29% (total payments rise by 2–40%)
Lyubchich and Gel (2017)	Daily number of flood-related home insurance claims in two Norwegian counties; 2002–2011	GARMA models (with alternating conditional expectations and regression trees for finding thresholds), and a new interval-based method for comparing tails in distributions	Daily precipitation	Predictions for 10-year periods in 2016–2095 based on seven combinations of climate models and climate scenarios. Projected increase is within 70–6,800%
Lyubchich et al. (2017)	Same as Soliman et al. (2015)	GARMA models for number of claims (with alternating conditional expectations and regression trees for finding thresholds); collective risk model for joint frequency-severity predictive modeling	Daily precipitation	Predictions for 2021–2030 based on two scenarios of CanESM2 climate model. Claims increase by 20–99%, 95th percentile of daily losses rises by 0–512%
Scheel and Hinnerichsen (2012) and Scheel et al. (2013)	Same as Haug et al. (2011), but focus on 319 municipalities in southern and central Norway	Bayesian Poisson hurdle model with Ising smoothed variable selection	Drainage runoff, precipitation on the previous day and early morning, and precipitation on the same day	No forecasts in the future
Soliman et al. (2015)	Property-level data on insurance claims and associated losses caused by weather and water damage in four Canadian cities; 2002–2011	GARMA models (with alternating conditional expectations for finding thresholds) and a new quantile-based method for comparing tails in distributions	Daily precipitation	Predictions for 10-year periods in 2021–2080 based on outputs from the CanESM2 climate model. Projected increase is within 4–62%
Spekkers et al. (2013)	Postal district-level daily data on number of policies, number of claims and associated losses due to water-related damages to private properties and content in the Netherlands; available since 1986 (property) or 1992 (content), but both analyzed for 2003–2009	Logistic regression	Maximum rainfall intensity	—
Spekkers et al. (2014)	Postal district-level daily data on number of policies, number of claims and associated losses due to water-related damages to private properties and content in the Netherlands (about 22% of all households in the country); 1998–2011	Decision (classification) tree	Maximum rainfall intensity	—
Spekkers et al. (2015)	Property-level data on claims in Rotterdam, the Netherlands, caused by water damage to building or content or both; January 2007–October 2013	Logistic regression	Maximum rainfall intensity	—

variables in $\mathbf{X}$, relatively equal importance of all the n observations, as well as uncorrelatedness, homoscedasticity, and normality of errors $\boldsymbol{\epsilon}$ (Chatterjee & Hadi, 2006). At the same time, distributions of losses are often heavily right-skewed; distributions of number of claims are discrete, often skewed to the right and inflated at zero. Hence, the assumption of normality is most often violated, and model (2) in its classical formulation cannot be applied to majority of insurance problems.

Generalized linear models (GLMs) help to overcome the violation of normality assumption by extending the applicability of model (2) to exponential-type distributions, such as Poisson, binomial, and gamma (Wood, 2006). In GLMs, distribution of Y_i belongs to a family of exponential distributions, and a smooth monotonic link function $g(\cdot)$ is applied to transform the response variable:

$$g(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}. \tag{3}$$

Canonical link functions are identity, ln, and inverse for normal, Poisson, and gamma distributions, respectively. After such transformation, however, model (3) still assumes linear relationships between each of the original variables in $\mathbf{X}$ and the transformed response. Model (3) is applicable when the link function successfully linearizes the relationship between the risk variable and a predictor. In other cases, especially if there are multiple predictors, additional work on respecifying the model may be required. For example, relationships between the risk variable and different predictors may require different linearizing transformations, the relationships may be nonmonotonic, and many of them may be thresholded (i.e., the effect of a covariate X is pronounced only when X takes on values from a certain range, such as the effect of daily precipitation amounts on the number of home insurance claims is not noticeable below certain precipitation threshold).

One way of capturing highly nonlinear relationships can be inclusion of additional transformed X-variables, such as power transformed or thresholded variables. However, adding tightly linked variables into the design matrix $\mathbf{X}$ may introduce multicollinearity and affect the inference. An alternative way of modeling nonlinearities is replacing the original variables with those individually transformed using smooth (nonparametric) functions, such as in a generalized additive model (GAM):

$$g(\boldsymbol{\mu}) = \mathbf{X}^*\boldsymbol{\beta}^* + f_1(X_1) + f_2(X_2) + f_3(X_3, X_4) + \ldots, \tag{4}$$

where Y_i still follows one of the exponential-family distributions; $\mathbf{X}^*$ and $\boldsymbol{\beta}^*$ are the remaining variables and associated coefficients in strictly parametric formulation; $f(\cdot)$ are smooth functions, often represented by regression splines (Wood, 2006). Model (4) can easily deal with deviations from normality and can accommodate nonlinearity and nonmonotonicity of individual relationships, however, the model still fails to address the issue of remaining dependencies in the errors (e.g., see Kohn, Schimek, & Smith, 2000).

Historical insurance and weather records often exhibit spatio-temporal autocorrelation that propagates into model residuals and renders the inference based on models (2), (3), and (4) unreliable. An extension of model (4) by Stasinopoulos and Rigby (2007) to $k = 1, 2, 3, 4$ parameters $\boldsymbol{\theta}_k$ of a distribution (not just the location parameter μ_i, but also scale σ_i, and shape—skewness and kurtosis; can be generalized for $k > 4$) allows fitting k individual models

$$g_k(\boldsymbol{\theta}_k) = h_k(\mathbf{X}_k, \boldsymbol{\beta}_k) + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}), \tag{5}$$

where $k = 1$ produces model for the mean; $h_k(\cdot)$ and $h_{jk}(\cdot)$ are nonlinear functions; $\boldsymbol{\beta}_k$ is a parameter vector of length J_k; $\mathbf{X}_k$ is an $n \times J_k$ design matrix; $\mathbf{x}_{jk}$ are vectors of length n. The additive terms in this generalized additive model for location, scale and shape (GAMLSS) provide a flexible framework to specify random effects and correlation structure as in mixed effects models (Zuur, Ieno, Walker, Saveliev, & Smith, 2009); see Table 3 by Stasinopoulos and Rigby (2007) for other possible specifications of the additive terms. Hence, models of the form (5) may be a good choice for insurance problems, because such models accommodate nonnormal distributions, possibly highly nonlinear relationships, and spatio-temporal dependencies in the data.

Another group of models, called generalized autoregressive moving average (GARMA), was developed by Benjamin, Rigby, and Stasinopoulos (2003) as a combination of GLM (3) with Box–Jenkins approach of modeling temporal dependence:

$$g(\boldsymbol{\mu}_t) = \eta_t = \mathbf{X}_t\boldsymbol{\beta} + \sum_{j=1}^{p} \phi_j\{g(y_{t-j}) - \mathbf{X}_{t-j}\boldsymbol{\beta}\} + \sum_{j=1}^{q} \theta_j\{g(y_{t-j}) - \eta_{t-j}\}, \tag{6}$$

where $t = 1, \ldots, n$ is the time index; φ_j, $j = 1, \ldots, p$, are autoregressive coefficients; θ_j, $j = 1, \ldots, q$, are moving average coefficients, and p and q are the autoregressive and moving average orders, respectively. Model (6) is efficient for dealing with individual time series; several studies mentioned in Table 3 used GARMA models with thresholded values of X-variables (e.g., precipitation amounts)—a transformation that could be approached using GAMLSS models (5) as well.

Notice that the issue of different reliability of individual measurements can be solved in models (2)–(6) by introducing predefined weights in the estimation process. An automatic tuning of weights for improved model performance is possible with a number of boosting algorithms, such as AdaBoost.M1 (Hastie, Tibshirani, & Friedman, 2009).

Overall, model (5) is a powerful and flexible choice for a variety of insurance problems, when data exhibit complex spatio-temporal dependence and do not adhere to commonly used distributions, such as normal or Poisson.

An alternative direction is to cast the regression framework in a Bayesian hierarchical setting, where parameters β follow some joint distribution with space–time dependence. The posterior analysis is then performed via Markov chain Monte Carlo (MCMC) procedures. One of the prominent examples of such approach in application to climate change and insurance risks is the Bayesian Poisson hurdle model, proposed by Scheel et al. (2013).

The challenges of using the above statistical models include the choice of predictors, their transformations, distribution of the response variable, and model specification, which can be attempted with a variety of criteria (e.g., Akaike and Bayesian information criteria—AIC and BIC) ubiquitous in statistical literature. Machine learning approaches offer more flexibility by relaxing the assumptions about distributions and forms of relationships, and providing automated solutions for learning meta-features from large amounts of data. At the same time, the large number of tuning parameters that inhere in a machine learning (especially in deep learning) method and their ability of changing the output or extending the computing time dramatically put out a warning for cautious implementation and interpretation of those methods.

5 | CONCLUSION

In the last few years we have been witnessing an ever growing body of scientific evidence that extreme weather events increase in frequency and intensity. This phenomenon has already implied a significant upward trend of insurance claims due to storms, hurricanes, floods, droughts, and other natural hazards. Insurance industry appears to be at the forefront of risk posed by adverse atmospheric events. And while nowadays it is broadly recognized that development of efficient risk mitigation and adaptation strategies is impossible without joint interdisciplinary efforts among actuaries, statisticians, atmospheric scientists, civil engineers and policy makers, such truly interdisciplinary initiatives are still relatively rare. Indeed, there yet exist a limited number of actuarial, statistical and climate studies that quantify and predict the impact of climate variability on the insurance industry, particularly, for the case of noncatastrophic, or low individual high cumulative impact events; while the statistical literature on the weather- and climate-induced risk in insurance is even scarcer. Furthermore, among the two major obstacles hindering efficient risk management due to natural hazards are lack of systematic records on insurance claims and limited understanding of uncertainties and their role in weather- and climate-induced risk mitigation. We believe that statistical sciences play a vital role in addressing those challenges and more generally in facilitating climate adaptation and insurance risk management.

The current paper is one of the first attempts to fill this gap and to offer a literature overview of the currently available methods for risk assessment due to natural hazards in agricultural and house insurance sectors.

ACKNOWLEDGMENTS

The research of V.L. was partially supported by the National Science Foundation of the USA grant #1739823. N.K.N. was supported by the Growing Forward II and Canadian Agriculture Partnerships (CAP) Canadian Federal Funding Programs (AAFC).

CONFLICT OF INTEREST

The authors have declared no conflicts of interest for this article.

RELATED WIREs ARTICLES

The durability of building materials under a changing climate
The resilience of integrated agricultural systems to climate change
Analyzing abrupt and nonlinear climate changes and their impacts

A review and classification of analytical methods for climate change adaptation
Micro-insurance for local adaptation

REFERENCES

AFSC. (2018). *Canada-Alberta AgriInsurance products for 2018 annual crops (Tech. Rep.)*. Lacombe, AB: Agriculture Financial Services Corporation Retrieved from https:// www.afsc.ca/doc.aspx?id=8067

Ahmed, O., & Serra, T. (2015). Economic analysis of the introduction of agricultural revenue insurance contracts in Spain using statistical copulas. *Agricultural Economics*, *46*(1), 69–79. https://doi.org/10.1111/agec.12141

Alexandridis, A. K., & Zapranis, A. D. (2013). *Weather derivatives: Modeling and pricing weather-related risk*. New York, NY: Springer-Verlag.

Benjamin, M. A., Rigby, R. A., & Stasinopoulos, D. M. (2003). Generalized autoregressive moving average models. *Journal of the American Statistical Association*, *98* (461), 214–223. https://doi.org/10.1198/016214503388619238

Black, E., Greatrex, H., Young, M., & Maidment, R. (2016). Incorporating satellite data into weather index insurance. *Bulletin of the American Meteorological Society*, *97*(10), ES203–ES206. https://doi.org/10.1175/BAMS-D-16-0148.1

Chatterjee, S., & Hadi, A. S. (2006). *Regression analysis by example*. Hoboken, NJ: John Wiley & Sons.

Cheng, C. S., Li, Q., Li, G., & Auld, H. (2012). Climate change and heavy rainfall-related water damage insurance claims and losses in Ontario, Canada. *Journal of Water Resource and Protection*, *4*, 49–62. https://doi.org/10.4236/jwarp.2012.42007

Choudhury, A., Jones, J., Okine, A., & Choudhury, R. (2016). Drought-triggered index insurance using cluster analysis of rainfall affected by climate change. *Journal of Insurance Issues*, *39*(2), 169–186.

Conradt, S., Finger, R., & Bokusheva, R. (2015). Tailored to the extremes: Quantile regression for index-based insurance contract design. *Agricultural Economics*, *46* (4), 537–547. https://doi.org/10.1111/agec.12180

Conradt, S., Finger, R., & Spörri, M. (2015). Flexible weather index-based insurance design. *Climate Risk Management*, *10*, 106–117. https://doi.org/10.1016/j.crm.2015.06.003

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. *Environmental Research Letters*, *13*(11), 114003. https://doi.org/10.1088/1748-9326/aae159

Curry, L., Weaver, A., & Wiebe, E. (2012). Determining the impact of climate change on insurance risk and the global community. Phase I: Key climate indicators (Tech. Rep.). Victoria, BC: American Academy of Actuaries' Property/Casualty Extreme Events Committee, CAS, CIA, and SOA. Retrieved from https://www.soa.org/ research-reports/2012/research-2012-climate-change-reports/

Dalhaus, T., & Finger, R. (2016). Can gridded precipitation data and phenological observations reduce basis risk of weather indexbased insurance? *Weather, Climate, and Society*, *8*(4), 409–419. https://doi.org/10.1175/WCAS-D-16-0020.1

Daron, J. D., & Stainforth, D. A. (2014). Assessing pricing assumptions for weather index insurance in a changing climate. *Climate Risk Management*, *1*, 76–91. https://doi.org/10.1016/j.crm.2014.01.001

de Leeuw, J., Vrieling, A., Shee, A., Atzberger, C., Hadgu, K. M., Biradar, C. M., … Turvey, C. (2014). The potential and uptake of remote sensing in insurance: A review. *Remote Sensing*, *6*(11), 10888–10912. https://doi.org/10.3390/rs61110888

Doms, J. (2017). Put, call or strangle? About the challenges in designing weather index insurances to hedge performance risk in agriculture. 57th Annual Conference, Weihenstephan, Germany, September 13-15, 2017 No. 261990. German Association of Agricultural Economists (GEWISOLA). Retrieved from https://ideas.repec.org/ p/ags/gewi17/261990.html

Donat, M. G., Leckebusch, G. C., Wild, S., & Ulbrich, U. (2010). Benefits and limitations of regional multi-model ensembles for storm loss estimations. *Climate Research*, *44*(2–3), 211–225. https://doi.org/10.3354/cr00891

Donat, M. G., Leckebusch, G. C., Wild, S., & Ulbrich, U. (2011). Future changes in European winter storm losses and extreme wind speeds inferred from GCM and RCM multi-model simulations. *Natural Hazards and Earth System Sciences*, *11*(5), 1351–1370. https://doi.org/10.5194/nhess-11-1351-2011

Environment Canada. (2017). *Top ten weather stories for 2012: Story four*. Ottawa: Environment of Canada. https:// ec.gc.ca/meteo-weather/default.asp?lang=En&n=70B4A3E9-1 (Online).

Erhardt, R. J. (2017). Climate, weather and environmental sources for actuaries (Tech. Rep.). Schaumburg, IL: Society of Actuaries. Retrieved from https://www.soa.org/ research-reports/2017/climate-weather-environmental-sources/.

Frees, E. W., Meyers, G., & Cummings, A. D. (2012). Predictive modeling of multi-peril homeowners insurance. *Variance*, *6*(1), 11–31.

Gerlt, S., Thompson, W., & Miller, D. J. (2014). Exploiting the relationship between farm-level yields and county-level yields for applied analysis. *Journal of Agricultural and Resource Economics*, *39*(2), 253–270.

Goodwin, B. K. (2015). Copula-based models of systemic risk in U.S. *American Journal of Agricultural Economics*, *97*(3), 879–896.

Grace, M. F., Klein, R. W., & Kleindorfer, P. R. (2004). Homeowners insurance with bundled catastrophe coverage. *Journal of Risk and Insurance*, *71*(3), 351–379. https://doi.org/10.1111/j.0022-4367.2004.00094.x

Hall, R. D. (2017, April/May 2017). Analyzing extreme weather. *The Actuary*.

Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). New York, NY: Springer. https://doi.org/10.1007/978-0-387-84858-7

Haug, O., Dimakos, X. K., Vardal, J. F., Aldrin, M., & Meze-Hausken, E. (2011). Future building water loss projections posed by climate change. *Scandinavian Actuarial Journal*, *1*, 1–20. https://doi.org/10.1080/03461230903266533

Held, H., Gerstengarbe, F.-W., Pardowitz, T., Pinto, J. G., Ulbrich, U., Born, K., … Burgho, O. (2013). Projections of global warming induced impacts on winterstorm losses in the German private household sector. *Climatic Change*, *121*(2), 195–207. https://doi.org/10.1007/s10584-013-0872-7

IFoA. (2017). Data science in insurance: Opportunities and risks for consumers—Policy briefing (Tech. Rep.). Institute and Faculty of Actuaries (IFoA). Retrieved from https://www.actuaries.org.uk/documents/policy-briefing-data-science-insurance-opportunities-and-risks-consumers.

Kamilaris, A., & Prenafeta-Boldu, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, *147*, 70–90. https://doi.org/10.1016/j.compag.2018.02.016

Klawa, M., & Ulbrich, U. (2003). A model for the estimation of storm losses and the identification of severe winter storms in Germany. *Natural Hazards and Earth System Science*, *3*(6), 725–732. https://doi.org/10.5194/nhess-3-725-2003

Kohn, R., Schimek, M. G., & Smith, M. (2000). Spline and kernel regression for dependent data. In M. G. Schimek (Ed.), *Smoothing and regression: Approaches, computation, and application* (pp. 135–158). New York, NY: John Wiley & Sons, Inc.. https://doi.org/10.1002/9781118150658.ch6

Kung, H.-Y., Kuo, T.-H., Chen, C.-H., & Tsai, P.-Y. (2016). Accuracy analysis mechanism for agriculture data using the ensemble neural network method. *Sustainability*, *8*(8), 735. https://doi.org/10.3390/su8080735

Kuwata, K., & Shibasaki, R. (2015). Estimating crop yields with deep learning and remotely sensed data. In 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), (pp. 858–861). Milan, Italy.

Kuwata, K., & Shibasaki, R. (2016). Estimating corn yield in the United States with MODIS EVI and machine learning methods. In ISPRS annals of photogrammetry, remote sensing & spatial information sciences, vol. 3 (8).

Leblois, A., & Quirion, P. (2013). Agricultural insurances based on meteorological indices: Realizations, methods and research challenges. *Meteorological Applications*, *20*(1), 1–9. https://doi.org/10.1002/met.303

Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), 2674. https://doi.org/10.3390/s18082674

Lyubchich, V., & Gel, Y. R. (2017). Can we weather proof our insurance? *Environmetrics*, *28*(2), e2433. https://doi.org/10.1002/env.2433

Lyubchich, V., Kilbourne, K. H., & Gel, Y. R. (2017). Where home insurance meets climate change: Making sense of climate risk, data uncertainty, and projections. Variance. Retrieved from https://www.variancejournal.org/articlespress/articles/Home-Lyubchich.pdf (in press).

Moriondo, M., Argenti, G., Ferrise, R., Dibari, C., Trombi, G., & Bindi, M. (2016). Heat stress and crop yields in the Mediterranean basin: Impact on expected insurance payouts. *Regional Environmental Change*, *16*(7), 1877–1890. https://doi.org/10.1007/s10113-015-0837-7

Newlands, N. K., Ghahari, A., Gel, Y. R., Lyubchich, V., & Mahdi, T. (2019). Deep learning for improved agricultural risk management. In Proceedings of the 52nd Hawaii international conference on system sciences (HICSS) (pp. 1033–1042). Maui, HI.

NOAA. (2018). U.S. Billion-dollar weather and climate disasters. National Centers for Environmental Information (NCEI). Retrieved from https://www.ncdc.noaa.gov/billions/.

Odening, M., & Shen, Z. (2014). Challenges of insuring weather risk in agriculture. *Agricultural Finance Review*, *74*(2), 188–199. https://doi.org/10.1108/AFR-11-2013-0039

Pacheco, J. M., Santos, F. C., & Levin, S. A. (2016). Evolutionary dynamics of collective index insurance. *Journal of Mathematical Biology*, *72*(4), 997–1010. https://doi.org/10.1007/s00285-015-0939-3

Pantazi, X. E., Moshou, D., Alexandridis, T., Whetton, R. L., & Mouazen, A. M. (2016). Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*, *121*, 57–65. https://doi.org/10.1016/j.compag.2015.11.018

Patel, K. (2018). Prolonged hot, dry conditions affect European crop prices. Retrieved from https://climate.nasa.gov/news/2806/prolonged-hot-dry-conditions-affect-european-crop-prices/.

Porth, L., & Tan, K. S. (2015). Agricultural insurance—More room to grow? *The Actuary Magazine*, *12*(2), 35–41.

Porth, L., Tan, K. S., & Zhu, W. (2016). Farm-level crop yield forecasting in the absence of farm-level data (Tech. Rep.). Schaumburg, IL: Society of Actuaries. Retrieved from https://www.soa.org/Files/Research/Projects/research-2016-farm-level-forecasting.pdf.

Pres, J. (2009). Measuring non-catastrophic weather risks for businesses. *The Geneva Papers on Risk and Insurance*, *34*(3), 425–439. https://doi.org/10.1057/gpp.2009.16

Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., & Meze-Hausken, E. (2013). 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. https://doi.org/10.1111/j.1467-9876.2012.01039.x

Scheel, I., & Hinnerichsen, M. (2012). The impact of climate change on precipitation-related insurance risk: A study of the effect of future scenarios on residential buildings in Norway. *The Geneva Papers on Risk and Insurance*, *37*(2), 365–376. https://doi.org/10.1057/gpp.2012.7

Shiraishi, H. (2016). Review of statistical actuarial risk modelling. *Cogent Mathematics*, *3*, 1123945. https://doi.org/10.1080/23311835.2015.1123945

Smith, A. B., & Katz, R. W. (2013). US billion-dollar weather and climate disasters: Data sources, trends, accuracy and biases. *Natural Hazards*, *67*(2), 387–410.

Smith, A. B., & Matthews, J. L. (2015). Quantifying uncertainty and variable sensitivity within the US billion-dollar weather and climate disaster cost estimates. *Natural Hazards*, *77*(3), 1829–1851. https://doi.org/10.1007/s11069-015-1678-x

Smith, V. H., & Glauber, J. W. (2012). Agricultural insurance in developed countries: Where have we been and where are we going? *Applied Economic Perspectives and Policy*, *34*(3), 363–390. https://doi.org/10.1093/aepp/pps029

Soliman, M., Lyubchich, V., Gel, Y. R., Naser, D., & Esterby, S. (2015). Evaluating the impact of climate change on dynamics of house insurance claims. In V. Lakshmanan, E. Gilleland, A. McGovern, & M. Tingley (Eds.), *Machine learning and data mining approaches to climate science* (pp. 175–183). Switzerland: Springer. https://doi.org/10.1007/978-3-319-17220-016

Spekkers, M. H., Clemens, F. H. L. R., & ten Veldhuis, J. A. E. (2015). On the occurrence of rainstorm damage based on home insurance and weather data. *Natural Hazards and Earth System Sciences*, *15*(2), 261–272. https://doi.org/10.5194/nhess-15-261-2015

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., & ten Veldhuis, J. A. E. (2013). A statistical analysis of insurance damage claims related to rainfall extremes. *Hydrology and Earth System Sciences*, *17*(3), 913–922. https://doi.org/10.5194/hess-17-913-2013

Spekkers, M. H., Kok, M., Clemens, F. H. L. R., & ten Veldhuis, J. A. E. (2014). Decision-tree analysis of factors influencing rainfall-related building structure and content damage. *Natural Hazards and Earth System Sciences*, *14*(9), 2531–2547. https://doi.org/10.5194/nhess-14-2531-2014

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (GAMLSS) in R. *Journal of Statistical Software*, *23*(7), 1–46.

Stulec, I. (2017). Effectiveness of weather derivatives as a risk management tool in food retail: The case of Croatia. *International Journal of Financial Studies*, *5*, 1–15. https://doi.org/10.3390/ijfs5010002

Tack, J. B., & Ubilava, D. (2015). Climate and agricultural risk: Measuring the effect of ENSO on US crop insurance. *Agricultural Economics*, *46*(2), 245–257. https://doi.org/10.1111/agec.12154

Toeglhofer, C., Mestel, R., & Prettenthaler, F. (2012). Weather value at risk: On the measurement of noncatastrophic weather risk. *Weather, Climate, and Society*, *4*(3), 190–199. https://doi.org/10.1175/WCAS-D-11-00062.1

Wood, S. N. (2006). *Generalized additive models: An introduction with R*. New York, NY: Chapman and Hall/CRC.

Woodward, J. D. (2016). Integrating high resolution soil data into federal crop insurance policy: Implications for policy and conservation. *Environmental Science & Policy*, *66*, 93–100. https://doi.org/10.1016/j.envsci.2016.08.011

Zuur, A., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). *Mixed effects models and extensions in ecology with R*. New York, NY: Springer. https://doi.org/10.1007/978-0-387-87458-6

How to cite this article: Lyubchich V, Newlands NK, Ghahari A, Mahdi T, Gel YR. Insurance risk assessment in the face of climate change: Integrating data science and statistics. *WIREs Comput Stat*. 2019;11:e1462. https://doi.org/10.1002/wics.1462