26th International Congress on Modelling and Simulation, Adelaide, SA, Australia, 30 Nov–4 Dec 2025
mssanz.org.au/modsim2025

Statistical Modelling of Property and Crop Losses

K Jason [a], N Raveendran [b], H Zhu [a], H Li [c], G Sofronov [a]

[a] *School of Mathematical and Physical Sciences, Macquarie University, NSW 2109, Australia*
[b] *School of Computer, Data and Mathematical Sciences, Western Sydney University, NSW 2116, Australia*
[c] *Centre for Actuarial Studies, Department of Economics, The University of Melbourne, VIC 3010, Australia*
Email: kristoforus.jason@mq.edu.au

Abstract: Natural disasters cause billions of dollars in property and crop damage annually in the United States alone. Modelling these losses is a core part of the insurance industry, as insurers need to understand the behaviour of the losses to set premiums that are both competitive and sustainable. Understanding the impact of each hazard can also help policymakers make informed decisions about allocating government resources toward natural disaster mitigation.

This study aims to model both property and crop losses simultaneously using existing R packages designed to handle multivariate responses, allowing us to capture the correlation between the two types of losses. Using SHELDUS loss data as the response variable, we incorporated various climate, geographical, and socioeconomic data in a GLM setting and identified key variables that are significant in explaining the losses. Preliminary exploratory analysis reveals a significant correlation between property and crop losses for hazards that cause the most damage. Using the `mcglm` package, we are able to obtain a joint model between the two types of losses that performs better than two separate models.

The loss dataset contains identical loss numbers due to different hazards in the same month and county, indicating potential inaccuracies in the data, possibly stemming from imputation, and thus limits the accuracy of our modelling. We also observe a surge in the number of events for certain hazards in some years, possibly due to the lack of reliable reports before the surge, which led us to focus our modelling on more recent years. Our analysis of the data reveals a correlation between different hazards, which can be a subject for future research considering compounding events.

Keywords: *GLM, multivariate response, mcglm, loss modelling, natural disaster.*

1 INTRODUCTION

Natural weather and climate disasters regularly cause billions of dollars in damages each year in the United States. In 2024 alone, the National Centers for Environmental Information (2025*a*) recorded 27 weather disaster events in the United States with losses of more than $1 billion dollars each, the most common suspect being severe storms and cyclones. Recent years have seen significantly more such costly events - the average number of billion-dollar events (CPI-adjusted) from 1980 to 2024 is 9 events, while the average for the last 5 years is 23 events. This shows a growing need to predict the impacts of these events in order to prepare for them and focus efforts toward damage mitigation.

In this study, we use the Spatial Hazard Events and Losses Database for the United States (SHELDUS), a database that records the occurrences of 18 types of natural disaster events in the United States down to the county level (2025). Among the variables recorded are the property and crop damages of each event. Although we have data from 1960 up to 2023, the decision was made to focus on modelling more recent years, from 2000 to 2023, as our exploratory analysis of the data reveals a sudden change in the number of events of certain hazards in the 1990s. A previous study by Raveendran et al. (2025), which uses another version of the same database, found the same shift in the data after a change point detection analysis. The availability of other variables serving as our explanatory variables also contributed to this decision. Ultimately, this allows us to obtain a more relevant model that reflects recent climate and economic conditions.

Generalised Linear Models (GLMs) are widely used in actuarial science to model losses which are often not normally distributed (see Haberman and Renshaw (1988)). We found a non-negligible positive correlation between the two types of damage in the data, which one can expect since these natural disaster events often affect large areas and thus affect both residential and agricultural assets. This motivates us to model the property and crop damages from the dataset simultaneously using existing GLM R packages designed to handle multivariate responses. Of course, each natural disaster affects property and crop damages differently, so the multivariate response models are separated by the type of hazard. Here, we use the ‘Flooding’ hazard as an example. We have decided to go with the `mcglm` package by Bonat (2018).

The loss data obtained from SHELDUS are combined with publicly available climate, geographical, and socioeconomic data. The climate data used in this study are obtained from NOAA’s National Centers for Environmental Information, organised by climate divisions - areas of the United States grouped based on climate similarity. We used the newer nClimDiv dataset, which is an improved dataset designed to address the issues the previous Drd964x dataset had (Vose et al. (2014)). Location data and population data are obtained from the US Census Bureau (2023, 2025). The economic data are obtained from the US Bureau of Economic Analysis (2025).

The rest of the article is structured as follows. Section 2 describes the different data sets that are considered in this study, the processing steps taken, and the results of exploratory data analysis. Section 3 outlines the methods used to model the losses. Section 4 presents the results and interpretations of the models.

2 DATA DESCRIPTION AND EXPLORATORY ANALYSIS

In this study, we are interested in analysing and modelling property and crop damage. For the explanatory variables, we consider climate, geographical, and socioeconomic factors obtained from publicly available sources. Table 1 contains all the variables considered for the models.

2.1 Data description

Spatial Hazard Events and Losses Database for the United States (SHELDUS) is a database maintained by the Center for Emergency Management and Homeland Security (CEMHS), Arizona State University (2025). The database covers natural events that occurred in the US and its territories from January 1960 to December 2023 (as of version 23.0). In this study, we used their direct loss data set. Each row contains information about the occurrence of a natural event in a county at a specified time. Among these variables, we chose the year, month, hazard type, FIPS (Federal Information Processing Standard) code, property damage, and crop damage. Both property damage and crop damage are adjusted for inflation to 2023. Note that SHELDUS refers to these damages as either “direct loss” or “damage”, both referring to the direct damages to property and crops, excluding any indirect and intangible costs. In this paper, we use “damage” and “loss” interchangeably, both meaning direct damages or direct losses only.

For the geographical data, we obtain information about each county from US Census Bureau (2023, 2024).

Table 1. List of variables considered for the models

Variable	Type	Available data
Year	Discrete	1960 - 2023
Month	Discrete	1960 - 2023
Hazard	Categorical	1960 - 2023
Crop Damage (adj. 2023)	Continuous	1960 - 2023
Property Damage (adj. 2023)	Continuous	1960 - 2023
FIPS / County Code	Categorical	All
Longitude	Continuous	All
Latitude	Continuous	All
Mean Elevation	Continuous	All
Area	Continuous	All
ACI Region	Categorical	All
Population	Continuous	2000 - 2023
Industry Total (Real GDP)	Continuous	2001 - 2023
Agriculture (Real GDP)	Continuous	2001 - 2023
Real Estate (Real GDP)	Continuous	2001 - 2023
Climate Division	Categorical	All
Precipitation	Continuous	1895 - 2023
Minimum Temperature (°F)	Continuous	1895 - 2023
Maximum Temperature (°F)	Continuous	1895 - 2023
Mean Temperature (°F)	Continuous	1895 - 2023
Palmer Drought Severity Index (PDSI)	Continuous	1895 - 2023

Note that the FIPS code is a unique identifier for these counties. For each county, we obtain the longitude, latitude, mean elevation, and the area of the county. We also assign each county to one of 6 regions, based on the classification used for the ACI (2018). The yearly population data for each county is also obtained from US Census Bureau (2010, 2020, 2025). We collect population data from 2000 to 2023.

For economic data, we collect real GDP data from US Bureau of Economic Analysis (2025), specifically the CAGDP9 data set. The data is available for each county from 2001 to 2023, and is separated by industry. We can use these economic data as the "exposure", as we expect the crop damage and property damage to scale with the size of the agriculture and real estate industry of each county. We consider the total GDP, the agriculture total, and the real estate total for the models. The GDP numbers in this data set are in chained 2017 dollars, which is a measure of real GDP. The problem with this data set is that a lot of the numbers are not disclosed to protect sensitive information. Thus, as a solution, we decided to take the average across all years to approximate the size of the industries of each county. We extrapolated this approximation to the year 2000, in order to be able to include that year into our modelling. Since we are using the average and only need to approximate industry sizes, we can retain the inflation adjustment as is.

Climate data is slightly different from the rest, as climates do not follow county lines. The data set obtained from National Centers for Environmental Information (2025*b*) divides the US region into climate divisions, where each division is supposed to have similar climate conditions. The newer data set that we use in this study (nClimDiv) is an improved version that addresses several problems the older one had (see Vose et al. (2014)). From this data set, we obtain monthly data for precipitation, minimum, maximum, and mean temperatures (in °F), and the Palmer Drought Severity Index (PDSI) (Palmer (1965)). We connect the climate divisions to the counties using the 'county-to-climdivs' matching provided in the same database. Climate data is available from 1895 to recent months, though we do not use the older years.

2.2 Exploratory analysis

The SHELDUS database has 18 different types of hazards. Our initial analysis is concerned with the actual number of natural events, rather than occurrences at the county level, which might be referring to the same event. To find this, we combined entries that have the same year and month, hazard, and state, while summing their property and crop damage. Figure 1 shows the annual count of the eight most common hazards.

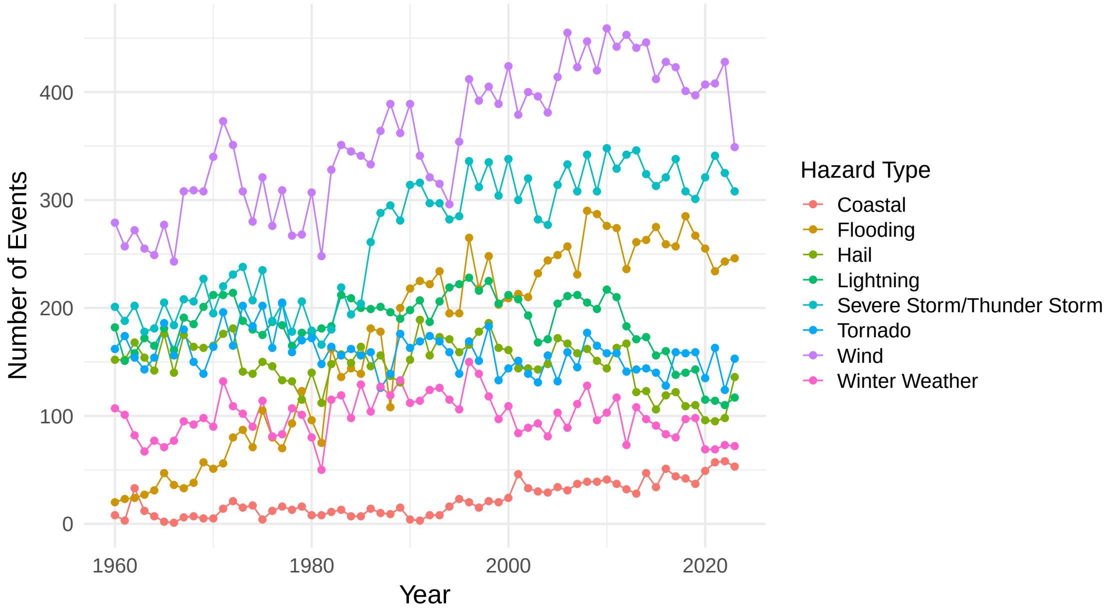

Figure 1. Annual count of the eight most common hazards in the US, from 1960 to 2023

We build a different model for each hazard, and for this paper, we use the hazard 'Flooding' as an example. Figure 2 illustrates the frequency of flood events in the US by state, from 1960 to 2023. Since we do have county-level data available, we perform our modelling on the county level in hopes of obtaining a more detailed result than a state-level model. Due to the availability of the variables in consideration, we decided to use data from year 2000 onwards, which will also better reflect the current climate and socioeconomic conditions.

Figure 3 is a scatterplot showing log-transformed crop damage versus property damage resulting from flood events from 2000 onwards, where each point corresponds to an entry in the SHELDUS data set. We can observe two distinct groups of points, one in the middle with positive crop and property damage, the other with zero crop damage or zero property damage. This study is interested in the first group, where we can observe a positive correlation of 0.74 between the two log-transformed damages. This relatively high correlation motivates us to perform a joint modelling.

3 MULTIVARIATE COVARIANCE GENERALIZED LINEAR MODELS

Bonat and Jørgensen (2016) introduced a general framework for non-normal multivariate data analysis called multivariate covariance generalized linear models (McGLMs). It can be viewed as an extension of the traditional GLM. The R package `mcglm` provides functions to fit and analyze McGLMs (Bonat (2018)).

Let $\mathbf{Y}_{N\times R} = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_R\}$ be the outcome matrix and let $\mathbf{M}_{N\times R} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_R\}$ be the corresponding matrix of expected values. Let $\boldsymbol{\Sigma}_r$ denote the $N \times N$ covariance matrix within the outcome r for $r = 1, \ldots, R$, and let $\boldsymbol{\Sigma}_b$ denote the $R \times R$ correlation matrix between the outcomes. The McGLMs are defined by:

$$\mathbb{E}(\mathbf{Y}) = \mathbf{M} = \left\{g_1^{-1}(\mathbf{X}_1\boldsymbol{\beta}_1), \ldots, g_R^{-1}(\mathbf{X}_R\boldsymbol{\beta}_R)\right\}, \tag{1}$$

$$\mathrm{Var}(\mathbf{Y}) = \mathbf{C} = \boldsymbol{\Sigma}_R \overset{G}{\bigotimes} \boldsymbol{\Sigma}_b, \tag{2}$$

where $\boldsymbol{\Sigma}_R \overset{G}{\bigotimes} \boldsymbol{\Sigma}_b = \mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1, \ldots, \tilde{\boldsymbol{\Sigma}}_R)\,(\boldsymbol{\Sigma}_b \otimes \mathbf{I})\,\mathrm{Bdiag}(\tilde{\boldsymbol{\Sigma}}_1^\top, \ldots, \tilde{\boldsymbol{\Sigma}}_R^\top)$ is the generalized Kronecker product. The matrix $\tilde{\boldsymbol{\Sigma}}_r$ denotes the lower triangular matrix of the Cholesky decomposition of $\boldsymbol{\Sigma}_r$. The operator Bdiag denotes a block diagonal matrix and $\mathbf{I}$ denotes an $N \times N$ identity matrix. The functions $g_r()$ are canonical link functions. Let $\mathbf{X}_r$ denote an $N \times k_r$ design matrix, and $\boldsymbol{\beta}_r$ denote a $k_r \times 1$ regression parameter vector.

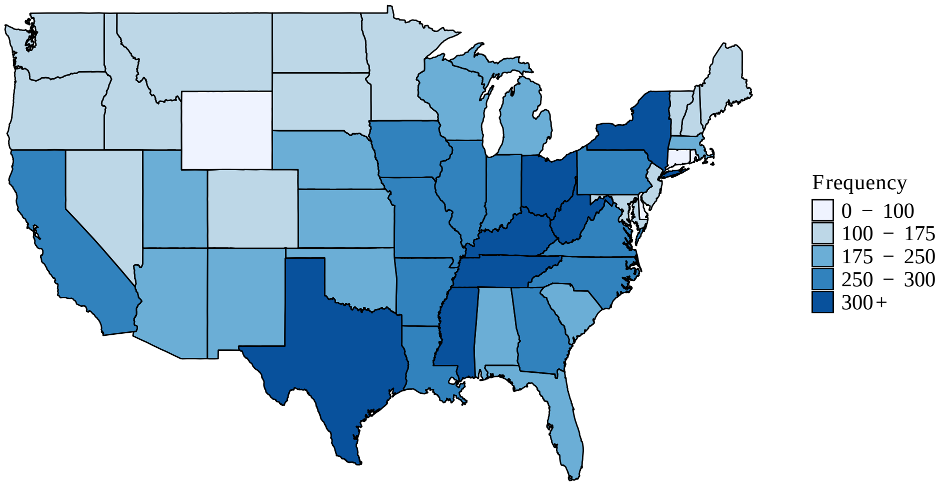

Figure 2. **U.S. Flood Events (1960 to 2023): Spatial Distribution and Frequency**

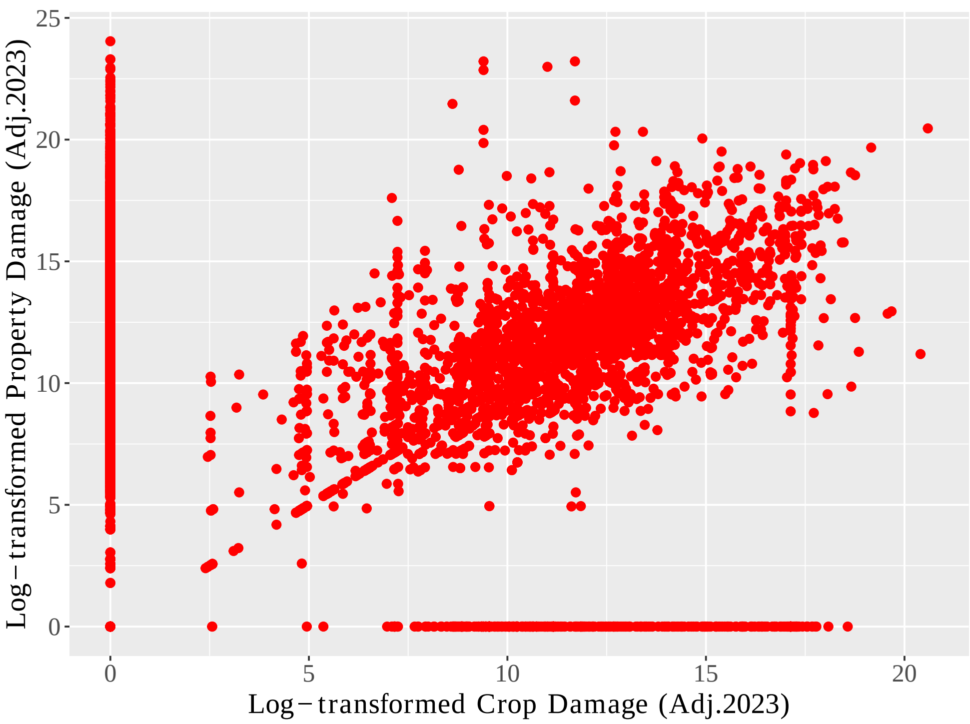

Figure 3. **Relationship between crop and property damages from flood events in the United States, 2000 to 2023**

Table 2. Crop damage model: estimated coefficients, standard errors, and p-values

Covariate	Estimate (SE)	p-value
(Intercept)	306.1322 (17.5131)	-
Region	-	0.0015
Year	-0.1502 (0.0087)	< 0.0001
Precipitation	0.1851 (0.0194)	< 0.0001
PDSI	0.1521 (0.0228)	< 0.0001
Log-transformed agriculture GDP	0.4827 (0.0333)	< 0.0001

In our case of continuous response variables, the covariance matrix within outcomes $\mathbf{\Sigma}_r$ is defined as:

$$\mathbf{\Sigma}_r = V(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}} \left(\mathbf{\Omega}(\boldsymbol{\tau}_r)\right) V(\boldsymbol{\mu}_r; p_r)^{\frac{1}{2}}, \tag{3}$$

where $V(\boldsymbol{\mu}_r; p) = \mathrm{diag}\{\vartheta(\boldsymbol{\mu}_r; p_r)\}$ is a diagonal matrix whose main entries are given by the variance function $\vartheta(\cdot; p_r)$ applied element-wise to the vector $\boldsymbol{\mu}_r$. The variance function for continuous outcomes is the power variance function $\vartheta(\cdot; p_r) = \mu_r^{p_r}$, which describes the Tweedie family of distributions. This includes the Gaussian ($p_r = 0$), Gamma ($p_r = 2$), and inverse Gaussian ($p_r = 3$) distributions.

The dispersion matrix $\mathbf{\Omega}(\boldsymbol{\tau}_r)$ describes the part of the covariance within outcomes that does not depend on the mean structure, modelled using a matrix linear predictor combined with a covariance link function:

$$h\left(\mathbf{\Omega}(\boldsymbol{\tau}_r)\right) = \tau_{r0} Z_{r0} + \tau_{r1} Z_{r1} + \cdots + \tau_{rD} Z_{rD}, \tag{4}$$

where $h()$ is the covariance link function, Z_{rd} are known matrices reflecting the covariance structure within the response variable r, and $\boldsymbol{\tau}_r = (\tau_{r0}, \cdots, \tau_{rD})$ is a $(D+1) \times 1$ parameter function. Bonat and Jørgensen proposed to model the spatial and temporal component into this covariance structure by using a linear combination of neighbourhood matrices. In their example (Example 3, Venezuelan rainfall data), spatial coordinates are used to build a Voronoi tessellation to help specify a neighbourhood structure, while temporal neighbours are naturally specified by the time structure. The interaction between space and time components is then described by the Kronecker product between their neighbourhood structures. In this paper, however, we will perform a simpler modelling that treats the space and time components as covariates. Future studies can incorporate these components into the covariance structure.

4 RESULTS AND CONCLUSIONS

In this paper, we performed modelling for the ‘Flooding’ hazard as an example. We modelled the spatial component using the ACI regions acting as a categorical variable and the time component using the year and/or month of the hazard events as variables. We encountered an error when using the log link function with the `mcglm` package, possibly due to the size of the dataset, so we opted to model the logarithm of the damages directly using the identity link function instead. After checking for collinearity between covariates, we selected significant variables using forward selection. We obtained the following models for crop damage ((5) and Table 2) and property damage ((6) and Table 3):

$$\log(\text{CropDamage}) \sim \text{Region} + \text{Year} + \text{Precipitation} + \text{PDSI} + \log(\text{AgricultureGDP}) \tag{5}$$

$$\log(\text{PropertyDamage}) \sim \text{Region} + \text{Year} + \text{Precipitation} + \text{PDSI} + \log(\text{RealEstateGDP}) \tag{6}$$

The goodness-of-fit test built into the package favours the joint model over modelling them separately. The multivariate model obtains a pseudo Akaike information criterion (pAIC) score of 30,580.02, while the single response variable models obtain a combined score of 33,035.04. This pAIC is a modified version of the AIC, see Bonat (2018) (Section 4) for more details. We note that the `mcglm` package does not provide the R-squared value as an output.

From the estimated coefficients, we can see that precipitation has a positive contribution towards both damages, which makes sense for the flood hazard. PDSI, which scales from -10 for extremely dry to +10 for extremely wet, also has a positive contribution. Agriculture and real estate GDP, which act as “exposures”, contribute positively towards crop damage and property damage, respectively. When comparing the magnitude of the coefficients from both models, precipitation appears to have a stronger effect on property damage, suggesting that urban infrastructure may be more vulnerable to flooding than agricultural systems.

Table 3. Property damage model: estimated coefficients, standard errors, and p-values

Covariate	Estimate (SE)	p-value
(Intercept)	298.3503 (17.2039)	-
Region	-	< 0.0001
Year	-0.1455 (0.0085)	< 0.0001
Precipitation	0.2818 (0.0191)	< 0.0001
PDSI	0.1478 (0.0224)	< 0.0001
Log-transformed real estate GDP	0.3198 (0.0238)	< 0.0001

The negative coefficient for 'Year' in both models suggests a downward trend in damages over time, which seems to contradict the current climate conditions. Considering the nature of the data set, this might suggest a declining insurance coverage for flooding over time, and thus fewer reported damages. This is consistent with an article from Moody's (2024), which noted that the National Flood Insurance Program (NFIP) sees a declining policy count, partly due to affordability issues. We also note that the SHELDUS data set itself contains inaccuracies due to imputations. In particular, when losses are reported as a range, SHELDUS takes the lower bound of the range as the losses. This limits the accuracy of any modelling that we perform.

While this study focuses on modelling losses for individual hazards, future research could explore the correlation between different hazard types, forming the possibility of compounding events. Additionally, incorporating the spatial and temporal components into the covariance structure could improve the accuracy of the model and provide more insight into where and when hazard events are more likely to occur.

REFERENCES

American Academy of Actuaries, Canadian Institute of Actuaries, Casualty Actuarial Society and Society of Actuaries (2018), Actuaries climate index executive summary.

Bonat WH (2018), Multiple response variables regression models in R: The mcglm package, *Journal of Statistical Software* **84**(4), 1–30.

Bonat WH and Jørgensen B (2016), Multivariate covariance generalized linear models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **65**(5), 649–675.

Center for Emergency Management and Homeland Security (CEMHS) (2025), Spatial hazard events and losses database for the United States (SHELDUS), version 23.0. Online database maintained by Arizona State University, Phoenix, AZ. Available at `https://hazards.asu.edu/sheldus`.

Haberman S and Renshaw AE (1988), Generalised linear models in actuarial work, Technical report, Institute and Faculty of Actuaries.

National Centers for Environmental Information (2025*a*), US billion-dollar weather and climate disasters.

National Centers for Environmental Information (2025*b*), US climate divisions – did you know?

Palmer WC (1965), Meteorological drought. US, *Weather Bureau Res. Paper* **45**, 1–58.

Raveendran N, Zhu H, Li H and Sofronov G (2025), Wildfire loss modeling: A flexible semiparametric approach, *North American Actuarial Journal* **29**(2), 329–344.

Saleh F (2024), How will the US flood insurance market evolve amidst rising risks and modeling advancements?

US Bureau of Economic Analysis (2025), CAGDP9 – real GDP by industry in chained dollars for counties and MSAs.

US Census Bureau (2010), County population totals: 2000-2009.

US Census Bureau (2020), County population totals: 2010-2019.

US Census Bureau (2023), Cartographic boundary shapefiles (GENZ2023).

US Census Bureau (2024), 2024 gazetteer files.

US Census Bureau (2025), County population totals: 2020-2024.

Vose RS, Applequist S, Squires M, Durre I, Menne MJ, Williams Jr CN, Fenimore C, Gleason K and Arndt D (2014), Improved historical temperature and precipitation time series for US climate divisions, *Journal of Applied Meteorology and Climatology* **53**(5), 1232–1251.

