Agricultural Insurance Loss and Relationships to Climate Across the Inland Pacific Northwest Region of the United States

A Dissertation
Presented in Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
with a
Major in Natural Resources
in the
College of Graduate Studies
University of Idaho
by

Erich Seamon

Major Professor: Paul E. Gessler, Ph.D.
Committee Members: John T. Abatzoglou, Ph.D.; Philip W. Mote, Ph.D.; Stephen S. Lee, Ph.D.

Department Administrator: Charles Goebel, Ph.D.

December 2019

Authorization to Submit Dissertation

This dissertation of Erich Seamon, submitted for the degree of Doctor of Philosophy with a Major in Natural Resources and titled "Agricultural Insurance Loss and Relationships to Climate Across the Inland Pacific Northwest Region of the United States," has been reviewed in final form. Permission, as indicated by the signatures and dates below, is now granted to submit final copies to the College of Graduate Studies for approval.

Major Professor _____ Date: _____
Paul E. Gessler, Ph.D.

Committee Members: _____ Date: _____
John T. Abatzoglou, Ph.D.

_____ Date: _____
Philip W. Mote, Ph.D.

_____ Date: _____
Stephen S. Lee, Ph.D.

Department Administrator: _____ Date: _____
Charles Goebel, Ph.D.

Abstract

Agricultural crop insurance is an important component for mitigating farm risk, particularly given the potential for unexpected climatic events (Christiansen et al., 1975; Diskin, 1997; Miranda & Glauber, 1997). Using a 2.8 million insurance claim dataset from the United States Department of Agriculture (USDA), this research study examined spatiotemporal variations of agricultural insurance loss across the 24-county region of the inland Pacific Northwest (iPNW) portion of the United States, from 2001 to 2015. For the prescribed time period, wheat was a dominant crop for the region, accounting for over $1.4 billion in insurance losses, with over $700 million resulting in claims due to drought. Principal components analysis showed distinct spatial and temporal differentiations in wheat insurance losses using the range of damage causes as factor loadings, with PC1 and PC2 accounting for 75% of total variance. Of particular note were the orthogonal relationships of county-level water availability damage causes (e.g. drought and heat vs. excessive moisture and cold wet winters), which aligned with regional climatic patterns. While both 2009 and 2015 were peak years for wheat/drought insurance loss, 2015 was the only year of the two that actually experienced regional drought conditions. Given the 2008/2009 recession economic impacts (Fan et al., 2015), this comparison may indicate the unique interaction of climate and economics and their impacts on climatically based damage cause insurance filings. Extending this iPNW analysis to evaluate quantitative climatic relationships to insurance loss, we developed a design matrix methodology to identify optimum temporal windows for climate variables by county, in relationship to wheat insurance loss due to drought. The results of our temporal window construction for water availability variables (precipitation, temperature, evapotranspiration, Palmer Drought Severity Index (PDSI)) identified spatial

patterns across the study area that aligned with regional climate patterns, particularly with regards to drought-prone counties of eastern Washington. Using these optimum time-lagged correlational relationships between insurance loss and individual climate variables, along with commodity pricing, we constructed a regression-based random forest model for insurance loss prediction, as well as to evaluate climatic feature importance. Our cross validated model results indicated that PDSI was the most important factor in predicting total seasonal wheat/drought insurance loss, with wheat pricing and potential evapotranspiration having noted contributions. Our overall regional model had a R^2 of .45, and a RMSE of ~$8.1 million. Model performance typically underestimated annual losses, with moderate spatial variability in terms of performance between counties.

Supporting our quantitative analysis of insurance loss and climate, we additionally constructed an open science-based framework for reproducibility (Flathers & Gessler, 2018; Fan et al., 2015; Dumontier & Wesley, 2018), applying our agricultural insurance loss and climate analysis as a case study example. Our case study framework implementation, which provided a set of dynamic analytics dashboards, code outputs, and reproducible research notebooks, identified several challenges that are critical for collaborative data intensive research, including: issues of data scaling, the importance of modular analytic code development, the challenges of team collaboration, data access and transformation and the difficulties regarding climatically gridded data, as well as institutional support for long term data resilience and viability.

Acknowledgements

I would like to initially thank my advisor, Paul Gessler, for the support and guidance he has provided throughout my dissertation writing, as well as committee members John Abatzoglou, Philip Mote, and Stephen Lee, who have inspired me to become a better scientist. In addition, I would like to recognize the Climate Impacts Research Consortium (http://pnwcirc.org), which funded this research, in concert with the National Oceanic and Atmospheric Administration's (NOAA) Regional Integrated Sciences and Assessment (RISA) program (award # NA15OAR4310145). I would finally like to thank my wife Masha and my daughter Natasha, for being so supportive and understanding.

Dedication

Dedicated to my daughter Natasha, who is the light of my life.

Table of Contents

List of Figures

List of Tables

CHAPTER 1: AGRICULTURAL INSURANCE LOSS ANALYSIS OF THE PACIFIC NORTHWEST, 2001-2015

1.1. Introduction

Agricultural systems are essential components to the Pacific Northwest (PNW) region of the United States (U.S.), encompassing the states of Idaho, Oregon, and Washington. As of 2015, agriculture accounted for over 500,000 jobs over these three states (Walker & Rahe 2015; Sandison 2015; Aviles et al., 2018). All three states consistently rank in the top five in terms of U.S. crop production for a range of agricultural commodities, including apples and wheat (Washington), potatoes and barley (Idaho), as well as hay, blackberries, and hazelnuts (Oregon) (United States Department of Agriculture [USDA] National Agricultural Statistics Service [NASS], 2016). In terms of agricultural exports, Washington ranked second behind California (2015), with Oregon placing eighth and Idaho, eleventh (USDA Economic Research Service [ERS], 2019).

Due to the considerable economic impact of agricultural commodity systems, as well as the potential negative implications due to unforeseen events, agricultural crop insurance has been an important component for mitigating risk (Christiansen et al., 1975; Diskin, 1997; Miranda and Glauber, 1997). In 1996 the USDA formed the Risk Management Agency (RMA), which works to increase the availability and effectiveness of federal crop insurance as a risk management tool. With the implementation of the Federal Crop Insurance Act (FCIA) and the USDA RMA, program improvements (i.e. providing direct payments to farmers, implementing subsidies) grew the level of program participation to over 90% of all U.S. farmed land by 1998. Crop insurance program efforts have had a dramatic impact on overall

farm management, including the reduction of income risk around crop production, increasing land values, increasing farm survivability rates, stabilizing cash flow, and liquidity improvement (Yu et al., 2018). By 2015, the USDA had insured over 114 million hectares of farmland, with an insurance liability net worth of $102 billion (USDA RMA, 2015). Given that the combined efforts of insurance protection as well as risk mitigation (related to agricultural practices) provide farmers a level of protection against unforeseen natural disasters and economic events, our research goal was to use exploratory data analysis techniques to examine spatiotemporal insurance loss claim variations across both the PNW, as well the iPNW subregion (Figure 1.1). We focused on losses due to weather and climate extremes, particularly those due to drought and heat, across agricultural commodities in the iPNW.

Weather and climate extremes, including those associated with climate change, have direct impacts on food security and resilience (Barrett et al., 2010; Gundersen et al., 2011). These interactions may vary due to a number of factors, including crop type, geographic location, and farming practices (Li et al., 2009). While previous studies have examined climate-yield relationships (Lobell & Burke 2010; Schlenker & Roberts 2009), there have been minimal analyses in examining climatic relationships related to crop insurance loss (Claassen et al., 2016; Schoengold, et al., 2014) Drought, in particular, plays an important role in the success or failure of many agricultural systems. Redmond (2002) conceptually defines drought as "insufficient water to meet needs", with a particular note of the varied relationships of supply and demand. Wilhite and Glantz (1985) describe drought broadly as a "deficiency of precipitation that results in water shortage for some activity or for some group" and emphasize the difficulties in having one overarching definition of drought, given its impacts

from an agricultural, climatological, meteorological, atmospheric, hydrologic, and water management perspective. Operationally, drought is often times quantified in terms of frequency, severity, intensity and duration, compared to a historical time frame, with human, biological, and climatological influences on both water supply and demand. Often times referred to as a "creeping phenomenon", the impacts of drought on society can persist for a number of years, dependent upon the level of vulnerability (Redmond, 2002). Agriculturally, drought often refers to a period with anomalously low soil moisture that substantially limits crop production (Mishra & Singh, 2010). Drought related impacts are evident in agricultural insurance loss claims, both nationally as well within the PNW. For example, drought conditions in 2015 resulted in agricultural insurance losses for PNW wheat alone totaling $183 million (USDA Risk Management Agency [RMA], 2015), with total financial losses for all commodities ranging between $633 million and $773 million (Washington State Department of Agriculture [WSDA], 2016). Adverse growing conditions over a season (such as during drought) can force farmers to consider additional risk management approaches that complement insurance mechanisms, including irrigation, selective crop abandonment, crop diversification, and unique crop rotation practices which may mitigate current and future losses and preserve long-term economic viability of cropping systems (Wallander et al., 2013; Yorgey & Kruger, 2017). For example, crop producers who utilize conservation tillage are often able to improve the capture and storage of soil moisture, which provides their crops an important buffer against drought impacts. By increasing the number of crop types as part of a rotation cycle, altering seeding dates, as well as using drought-sensitive breeds, farmers can retain more available soil moisture (reducing long term drawdown), while maximizing production and sales by spreading risk across a larger set of commodities

(Antle & Capalbo, 2010). From an adaptive perspective, the economic implications of more severe drought conditions, as well as a change in seasonal ranges of climatic conditions, may encourage farmers to consider alternative crop systems that are more economically viable, such as profitable niche fruit commodities. In total, these added risk management efforts, in combination with crop insurance, provide farmers with a diversified ability to mitigate potential financial loss in the face of changing economic and climatic conditions.

1.2. Data and Methods

The USDA's data archive of agricultural insurance claim records for the PNW, from 1989 to 2015, is the primary dataset for this analysis (http://usda.gov/rma). Insurance claims were provided at monthly temporal and county level spatial scale. Each insurance record represents a claim filed by a farmer, containing the dollar amount of the insured loss, the commodity type related to the loss (e.g. wheat, barley, canola), the acreage for the loss, and most notably, a cause for the crop damage (e.g. heat, drought, hail, decline in price, failure of irrigation supply). The extent of this data archive is considerable: for example, from 1989 to 2015, the USDA's crop insurance data collection for the United States totals approximately 2.8 million claims, with ~35,000 claims originating in the Pacific Northwest (Idaho, Oregon, and Washington) for over 35 different commodities, across 30 different damage causes (Appendix A).

We constructed a basic three step exploratory data analysis (EDA) methodology that allowed us to systematically examine commodity-specific insurance loss across damage causes in the PNW. EDA is an established approach (Tukey, 1977; Cleveland, 1993) to informally examine and refine large datasets, in preparation for more formal statistical and inferential

statistical analysis. Numerous techniques can be used as part of this process, including; evaluating data for mistakes and potential limitations, the checking of particular assumptions, determining which model(s) may be appropriate for future data evaluation, and examining variable relationships and data refinements based on spatial and temporal scales (Behrens, 1997). Given our research goal to examine how spatiotemporal variation of agricultural insurance loss reflects known climatological and economic historical trends, the results of these EDA steps not only allowed us to narrow our factorial analysis by geography, time, commodity, and damage cause, but also to compare how water scarcity-specific damage causes (drought and heat) varied based on these refined factors. Our EDA refinement process is described as follows:

1. We initially performed a full examination of insurance loss across all commodities and damage causes, for the entire PNW region, from 1989 to 2015. As part of this step, we aggregated the data by county, commodity, year, and damage cause. An initial data review indicated that approximately 83% of insurance loss for the region occurred from after 2000, which comports with farm bill policy incentives implemented in 1998, increasing crop insurance participation (acres) to over 90% (USDA, 2015). Spatially, over 75% of insurance losses occurred in the iPNW, with wheat losses being the overwhelming dominant commodity. As such, we limited our time frame of insurance loss examination to 2001 to 2015, and narrowed our study area region to the 24-county region of the IPNW (Figure 1.2). This reduction of data by year additionally helped to resolve missing data issues in some counties that had no insurance claims, and thus no revenue loss. Given the spatial diversity in terms of cropping systems across all three states, the IPNW area provided a more

homogenous, well distributed dryland farming region to allow us to explore spatial and temporal variations, while maintaining a fairly consistent county level claim total across the region as a whole. Additionally, by narrowing our region to only the IPNW, we eliminated counties where little or no insurance claims were filed, primarily due to landscape, urbanization, or profitability constraints.

2. We used principal components analysis (PCA) to identify commonalities in insurance claims across years, counties, commodities, and damage claims in the iPNW. PCA is a data dimensionality reduction technique which computes a new set of variables by maximizing the variance of all input variables, and then examines the linear combinations of said variables in orthogonal space (Tukey, 1977; Jolliffe, 2002). In our agricultural insurance loss analysis, we transformed our damage cause factorial observations, constructing individual variables for each damage cause insurance loss total ($), by county, commodity and year. Using this approach, we were able to create a set of input variables for our PCA, to examine how damage cause factors were associated, as well as how counties and years were aligned to these individual factor loading vectors. In order to evaluate how PCA variables group together, we apply a kmeans algorithm (Ding, 2004) to estimate optimal clusters for both county and year.

3. From the results of our EDA and PCA, we limited our commodity analysis to wheat, and narrowed our set of damage cause claims to areas of water scarcity (drought and heat). We then examined wheat loss for the region, exploring these water scarcity temporal and spatial relationships on an annual basis. In addition, we compared

insurance loss with overall wheat production across the 24-county study area, as well as annually (from 2001 to 2015).

1.3. Results

PNW insurance claims totaled over 28,000 from 2001 to 2015, for all commodities, with overall insured losses of $2.9 billion. Wheat, the dominant commodity for insurance claims in the three-state region, accounted for approximately 20,600 filings, with total losses of $1.45 billion for the same time period. (Figure 1.3). Cherries and apples were a distant second and third in terms of overall losses, each with approximately $180 million, with potatoes and peas adding a minimal contribution to the overall total. Narrowing our analysis to the iPNW, we see that insurance losses there made up approximately 72% of the total amount of loss for PNW as a whole. Wheat was similarly the predominant commodity incurring insurance loss for the iPNW, with over $1.2 billion in claims, with apples coming in a distant second, at $52 million. In term of damage cause, drought resulted in the largest amount of insurance loss for the PNW overall, at over $700 million, with decline in price ($340 million) and heat ($270 million) coming in second and third, respectively (Figure 1.4). Similarly, the leading damage causes for the iPNW were drought and heat, which combined to account for approximately $800 million in losses from 2001 to 2015.

In order to address our research question around spatial and temporal variations of insurance loss related to water availability, annual wheat losses due to drought, heat, and excessive moisture for the iPNW were analyzed for each year in the period from 2001 to 2015. The results show that the year-to-year variation of losses are dominated by drought, with peak years occurring in 2009 and 2015. In contrast, 2011 had almost no drought or heat insurance

losses, with excessive moisture and rain being the dominant damage cause factors (Figure 1.5). This annual variability aligns with historical climatological variations: 2011 was a particularly wet year for the PNW (NOAA National Centers for Environmental Information, 2011), while in 2015, the PNW experienced a significant drought (Fosu et al., 2016; Mote et al., 2016). If decline in price is incorporated into this annual view, we see the large majority of these claims occurring during the 2009 year, given wheat prices declines from ~$430 /metric ton to $220 /metric ton. Incorporating wheat production into this analysis, we see an inverse relationship, with the lowest levels of production occurring in years with the highest levels of drought/heat insurance loss (Appendix A)

Spatially, we initially analyzed county level wheat production from 2001 to 2015 (Figure 1.6), in relationship to drought/heat insurance losses for the same time period (Figure 1.7). While total 2001 to 2015 wheat production was highest in Whitman County, WA ($384 million), wheat insurance loss due to drought and heat were highest along the northeastern portion of the Oregon high desert (Umatilla county, Oregon at $118 million and Morrow county, Oregon at $97 million). From a percentage breakdown, over 50% of all damage cause losses in Umatilla were a result of drought/heat, with an over 30% ratio in Adams and Lincoln counties, Washington. If we more specifically examine spatial differences in wheat drought/heat insurance loss by year, we see notably different patterns of loss concentrations between 2011 and 2015. For 2011, the region's few drought and heat claims were concentrated in the eastern portion of the region, primarily in Idaho, with losses in the highly productive Palouse region of eastern Washington being relatively low (Figure 1.8). In contrast, 2015 wheat losses due to drought and heat were concentrated in the upper portion of the Washington Palouse region (Whitman, Lincoln, Adams, and Douglas counties), with

additional loss concentrations falling along the Columbia river valley and in the western portion of the Palouse (Figure 1.9). In order to better understand the factorial relationships of damage causes, two principal components analyses were run for the iPNW region to explore 1) spatial (county) as well as 2) temporal (year) variation. Both PC analyses used damage causes as the factor loadings, with all data scaled by the unit variance. Additionally we used singular value decomposition (SVD), a form of matrix factorization which is considered a superior method for PCA computation (Tanwar et al., 2018).

The results of this analysis indicate that approximately 75% of total variance of insurance loss by county level damage cause can be attributed to the first two principal components, with water scarcity (drought/heat/fire) damage causes having a negative coordinate alignment in terms of the first principal component (PC1) vector loading directions (Figure 1.10). Examining PC loadings by county (Figures 1.11 and 1.12), we see a clear alignment of water scarcity damage causes in highly productive wheat counties (Umatilla county, OR, Lincoln and Whitman counties, WA), with orthogonal damage causes (excessive moisture/freeze/frost) aligning with counties that are typically in highly productive fruit production regions (e.g. Grant and Benton counties, WA). Applying a kmeans clustering algorithm with an elbow cluster optimization selection method, we identified four key clusters in the two-dimensional PCA space, that additionally support the differentiation of water scarcity PC1 loadings from PC2 water excess.

When PCA was run using year as the independent factor (2001 to 2015), we see similar water scarcity/water excess damage cause loading groupings, with PC1 and PC2 resulting in approximately 64% of total variance (Figure 1.13). Kmeans clustering identified three key

groupings of years: most notably, 2009, 2014, and 2015 were distinct clusters along damage cause groupings for drought, fire, and heat. When PC1 and PC2 loadings are linearly compared by year (Figure 1.14), we see similar distinct divergences/opposing alignments between PC1 and PC2, in 2009, 2014, and 2015. Such divergence reinforces the clustering results for our yearly based PCA, which groups 2009, 2014, and 2015. Additional PCA for differing combinations (e.g. commodity) are included in Appendix B which provides an extended comparison of how insurance loss may vary across differing factorial combinations.

1.4. Conclusions

Given our research focus of using EDA to examine iPNW spatiotemporal variations of insurance loss in relationship to climatic damage causes, our results identified several unique spatial and temporal patterns that appear to align with historical climatological patterns. In addition, said patterns provide an important perspective on climate variability, economics, and the sensitivities of agricultural systems. Of particular interest were the differences in iPNW wheat insurance loss, comparing 2009, 2011, and 2015, in terms of the drought, heat, excessive moisture, and decline in price total losses. While 2009 and 2015 have large dollar losses in terms of drought and heat, 2009 additionally had extremely large values with regards to decline in price insurance claims, associated with a reduction in wheat prices (Fan et al., 2015). While increased drought and heat losses in 2015 align well with regional drought conditions (Marlier et al., 2017), increased drought and heat claims for 2009 seem to conflict with comparable climate conditions for that year, which indicates that the PNW was not in a period of drought (Shukla et al., 2015). These insurance loss comparisons between 2009 and 2015 suggest that, in compromised economic conditions (e.g. price decline), claims

due to climatic damage causes may increase, even though actual climatic conditions do not warrant such increases (Miranda & Vedenov, 2001, Botzen & Van Den Bergh, 2008). This may also indicate that particular commodity-specific thresholds exist where economic factors dominate over climatic impacts, resulting in a broad distribution of claim loss across a range of damage causes. 2011 losses were interestingly juxtaposed to 2009 and 2015, with very little drought or heat insurance claims, but with the largest amount of excessive moisture filings of any year in the period of analysis. We also saw an inverse relationship between annual wheat production and drought/heat insurance loss, with 2009 and 2015 being the only years in this time period where losses were higher than production. Work by Quiggin et al. (1993), Miranda et al. (1997), and Glauber (2004) all reference these relationships of insurance loss with overall crop production, supporting this inverse relationship scenario. Spatial variations of wheat insurance losses due to drought and heat provide an additional perspective in terms of locational sensitivities to climate. With variations of phenology, claim frequency, regional crop development, irrigation, and cropping practices, commodity-based insurance claim analysis for agriculturally homogeneous regions may provide the best framework for delineating differences in claim/loss variation, based on time and the cause of damage. These distinct differences in annual variation as well as commodity/damage cause indicate the sensitive aspects of insurance claim loss, which may serve as a more effective barometer in gauging climatic influences as compared to crop yield production given varying aspects of pricing, land values, equipment, input price changes, and climate (Fan et al., 2015). Our results highlight that insurance losses likely integrate aspects of climate and economic impact together (e.g. comparisons of 2009 and 2015 damage causes), given that farmer decisions regarding whether to file a loss claim or not typically take into account

these two factors jointly. The results of this work additionally highlight several areas of potential future research, particularly around understanding the interactions between insurance loss, conservation practices, economic factors, climate influences, and policy effects, as well as regional differences/similarities of damage cause influences across a range of commodities other than wheat. Under changing climate and conservation practice conditions, there may be situations where crop insurance risk management may incentivize, or disincentivize, farm practices that reduce agricultural climate change impacts, given their individualized economic implications. Additionally, this work may assist future research in identifying the financial impacts of a changing climate on insurance loss, over time and differing geographies.

1.5. References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. International Journal of Climatology, 33(1), 121–131.

Abatzoglou, J. T., & Brown, T. J. (2012). A comparison of statistical downscaling methods suited for wildfire applications. International Journal of Climatology, 32(5), 772–780. http://doi.org/10.1002/joc.2312

Abatzoglou, J. T., Rupp, D. E., & Mote, P. W. (2014). Seasonal climate variability and change in the pacific northwest of the united states. Journal of Climate, 27(5), 2125–2142. http://doi.org/10.1175/JCLI-D-13-00218.1

Alston, J. M., Beddow, J. M., & Pardey, P. G. (2009). Agricultural Research, Productivity, and Food Prices in the Long Run. Science, 325(September), 4–5.

Antle, J. M., & Capalbo, S. M. (2010). Adaptation of agricultural and food systems to climate change: An economic and policy perspective. Applied Economic Perspectives and Policy, 32(3), 386–416. http://doi.org/10.1093/aepp/ppq015

Aviles, G., Lee, D., Robinchtein, A. & Zoe Zadworny, Z. (2018). 2016 Agricultural Workforce Report. Washington State Employment Security Department Workforce Information and Technology Services.

Barrett, C. B. (2010). Measuring Food Insecurity. Science, 327(February), 825–828. http://doi.org/10.1126/science.1182768

Beguería, S., Vicente-Serrano, S. M., Reig, F., & Latorre, B. (2014). Standardized precipitation evapotranspiration index (SPEI) revisited: Parameter fitting, evapotranspiration models, tools, datasets and drought monitoring. International Journal of Climatology, 34(10), 3001–3023. http://doi.org/10.1002/joc.3887

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis in: Psychological Methods Vol.2. Psychological Methods, 2(2), 131–160.

Botzen, W. J. W., & Van Den Bergh (2008). Insurance against climate change and flooding in the Netherlands: Present, future, and comparison with other countries. Risk Analysis, 28(2), 413–426. https://doi.org/10.1111/j.1539-6924.2008.01035.x

Christensen, L. R. (1975). Concepts and Measurement of Agricultural Productivity. American Journal of Agricultural Economics, 57(5), 910. http://doi.org/10.2307/1239102

Claassen, R., Langpap, C., & Wu, J. (2016). Impacts of Federal Crop Insurance on Land Use and Environmental Quality. American Journal of Agricultural Economics, (June), aaw075. https://doi.org/10.1093/ajae/aaw075

Cleveland, W. S. (1993). Visualizing data. Murray Hill, N.J. : [Summit, N.J.]: At & T Bell Laboratories ; [Published by Hobart Press].

Crop Insurance: A Look Back: Crop Insurance & Crop Protection from Federal Crop Insurance Programs (2014). National Crop Insurance Services (NCIS). http://cropinsuranceinamerica.com/about-crop-insurance/history/

Deschênes, O., & Greenstone, M. (2007). The economic impacts of climate change: evidence from agricultural output and random variations in weather. The American Economic Review, 97(01), 354–385. http://doi.org/10.2307/3078074

Ding, C. (2004). K -means Clustering via Principal Component Analysis.

Diskin, P. (1997). Agricultural Productivity Indicators Measurement Guide. Food and Nutrition Technical Assistance Project (FANTA), (December).

Dumontier, M., & Wesley, K. (2018). Advancing discovery science with fair data stewardship: Findable, accessible, interoperable, reusable. Serials Librarian, 74(1–4), 39–48. https://doi.org/10.1080/0361526X.2018.1443651

Fan, M., Pena, A. & Perloff, J.M. (2015). Effects of the Great Recession on the U.S. Agricultural Labor Market. IRLE Working Paper No. 104-15. http://irle.berkeley.edu/workingpapers/104-15.pdf

Flathers, E. & Gessler, P.E. (2018). Building an Open Science Framework to Model Soil Organic Carbon. J. Environ. Qual. 47:726-734. doi:10.2134/jeq2017.08.0318

Food, Conservation, and Energy Act of 2008. (2008). Pub.Law 110–234, H.R. 2419, 122 Stat. 923, enacted May 22, 2008.

Fosu, B., Wang, S., & Yoon, J. (2016). The 2014/2015 snowpack drought in Washington state and its climate forcing. [in "Explaining Extremes of 2015 from a Climate Perspective"]. Bull. Amer. Meteor. Soc., 97 (12), S14–S18, doi:10.1175/BAMS-D-16-0149.

Glauber, J.W. (2004). Crop Insurance Reconsidered, American Journal of Agricultural Economics, Volume 86, Issue 5, December, Pages 1179–1195, https://doi.org/10.1111/j.0002-9092.2004.00663.x

Gundersen, C., Kreider, B., & Pepper, J. (2011). The economics of food insecurity in the United States. Applied Economic Perspectives and Policy, 33(3), 281–303. http://doi.org/10.1093/aepp/ppr022

Jolliffe, I. T. (2002). Principal Component Analysis. New York: Springer-Verlag. doi:10.1007/b98835

Kucharik, C. J., & Serbin, S. P. (2008). Impacts of recent climate change on Wisconsin corn and soybean yield trends. Environmental Research Letters, 3(3). http://doi.org/10.1088/1748-9326/3/3/034003

Li, Y., Ye, W., Wang, M., & Yan, X. (2009). Climate change and drought: a risk assessment of crop-yield impacts. Climate Research, 39(June), 31–46. http://doi.org/10.3354/cr00797

Lobell, D. B., Burke, M. B., Tebaldi, C., Mastrandrea, M. D., Falcon, W. P., & Naylor, R. L. (2008). Prioritizing Climate Change Adaptation Needs for Food Security in 2030. Science, 319(5863), 607–610. http://doi.org/10.1126/science.1152339

Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. Agricultural and Forest Meteorology, 150(11), 1443–1452. http://doi.org/10.1016/j.agrformet.2010.07.008

Lobell, D. B., Schlenker, W., Costa-Robert, J. (2011). Climate trends and global crop production since 1980. Science, 333(2011), 616–620. http://doi.org/10.1126/science.1204531

Marlier, M. E., Xiao, M., Engel, R., Livneh, B., Abatzoglou, J. T., & Lettenmaier, D. P. (2017). The 2015 drought in Washington State: a harbinger of things to come? Environmental Research Letters, 12(11), 114008. https://doi.org/10.1088/1748-9326/aa8fde

McCarl, B. A., Villavicencio, X., & Wu, X. (2008). Climate change and future analysis: Is stationarity dying? American Journal of Agricultural Economics, 90(5), 1241–1247. http://doi.org/10.1111/j.1467-8276.2008.01211.x

Milly, P. C. D., & Dunne, K. A. (2016). Potential evapotranspiration and continental drying. Nature Climate Change, 6(10), 946–949. http://doi.org/10.1038/nclimate3046

Miranda, M. J., & Glauber, J. W. (1997). Systemic Risk, Reinsurance, and the Failure of Crop Insurance Markets. American Journal of Agricultural Economics, 79(1), 206–215. http://doi.org/10.2307/1243954

Miranda, M., & Vedenov, D. V. (2001). Innovations in agricultural and natural disaster insurance. American Journal of Agricultural Economics, 83(3), 650-655.

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. Journal of Hydrology, 391(1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Mote, P. W., Rupp, D. E., Li, S., Sharp, D. J., Otto, F., Uhe, . . . Allen, M. R. (2016). 2015 snowpack in the western United States. Geophysical Research Letters, 1–9. http://doi.org/10.1002/2016GL069965

NOAA National Centers for Environmental Information. State of the Climate: Drought for March 2011. (2011). from https://www.ncdc.noaa.gov/sotc/drought/201103.

Quiggin, J. C., Karagiannis, G., & Stanton, J. (1993). Crop insurance and crop production: an empirical study of moral hazard and adverse selection. Australian Journal of Agricultural Economics, 37(429-2016-29192), 95.

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical Linear Models: Applications and data analysis methods (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Redmond, K. (2002). the Depiction of Drought. Bams, (December), 1143–1147.

Sandison, D. I. (2015). 2015 Drought and Agriculture, 495 (February). Retrieved from https://agr.wa.gov/FP/Pubs/docs/495-2015DroughtReport.pdf

Schoengold, K., Ding, Y., & Headlee, R. (2014). The impact of AD HOC disaster and crop insurance programs on the use of risk-reducing conservation tillage practices. American Journal of Agricultural Economics, 97(3), 897–919. https://doi.org/10.1093/ajae/aau073

Seltman, H. J. (1997). Experimental design and analysis. Experimental Design and Analysis, 7–10. http://doi.org/citeulike-article-id:9144499

Shukla, S., Safeeq, M., Aghakouchak, A., Guan, K., & Funk, C. (2015). Temperature impacts on the water year 2014 drought in California. Geophysical Research Letters, 42(11), 4384–4393. https://doi.org/10.1002/2015GL063666

Sorte, B., & Rahe, M. (2015). Oregon Agriculture, Food and Fiber: An Economic Analysis. Oregon Department of Agriculture, Oregon State University Extension Service, Rural Studies Program (December).

Sparks, B. (2013). Apple Crop Estimate Unveiled (2013). Growing Produce, August 2013. https://www.growingproduce.com/fruits/2013-apple-crop-estimate-unveiled/

Tanwar, S., Ramani, T., & Tyagi, S. (2018). Dimensionality Reduction Using PCA and SVD in Big Data: A Comparative Case Study. In Z. Patel & S. Gupta (Eds.), Future Internet Technologies and Trends (pp. 116–125). Cham: Springer International Publishing.

Thornton, P. K., Jones, P. G., Alagarswamy, G., & Andresen, J. (2009). Spatial variation of crop yield response to climate change in East Africa. Global Environmental Change, 19(1), 54–65. http://doi.org/10.1016/j.gloenvcha.2008.08.005

Tukey, J. W. (1977). Exploratory data analysis. Reading, MA: Addison-Wesley

United States Crop Insurance Program (2014). United States Department of Agriculture (USDA) Risk Management Agency. http://www.usda.gov/rma

U.S. Department of Agriculture, National Agricultural Statistics Service. (2016). 2016 Census of Agriculture, Vol. 1: Part 51, Chapter 2, AC97-A-51, United States Summary and State Data.

Walker, K., & Rahe, M. (2015). Oregon Agriculture, Food and Fiber: An Economic Analysis, (December).

Wallander, S., Aillery, M., Hellerstein, D., & Hand, M. (2013). The Role of Conservation Programs in Drought Risk Adaptation. Economic Research Service, April (148), 1–68. Retrieved from http://ers.usda.gov/media/1094684/err-148-summary.pdf

Wheat, D. (2017). "Second-largest crop pushes many apple prices lower" Capital Press, May 2017.

Wilhite, D. A. & Glantz, M. H. (1985). Understanding the drought phenomenon: The role of definitions. Water International 10: 111–20.

Yu, J., & Sumner, D. A. (2018). Effects of subsidized crop insurance on crop choices. Agricultural Economics (United Kingdom), 49(4), 533–545. https://doi.org/10.1111/agec.12434

Wilhite, D. A. (1992). Drought. Encyclopedia of Earth System Science, Vol. 2, pp. 81–92, San Diego, CA: Academic Press.

Yorgey, G. & Kruger, C. E. (2017). Advances in Dryland Farming in the Inland Pacific Northwest. Washington State University Extension Publications. Retrieved from https://books.google.com/books?id=vZnEswEACAAJ

Figure 1.1. 24-county inland Pacific Northwest (iNPW) study area, which includes counties from Washington, Idaho, and Oregon. Additionally noted: on the inset map in upper left depicts the three main agricultural regions in the Pacific Northwest.

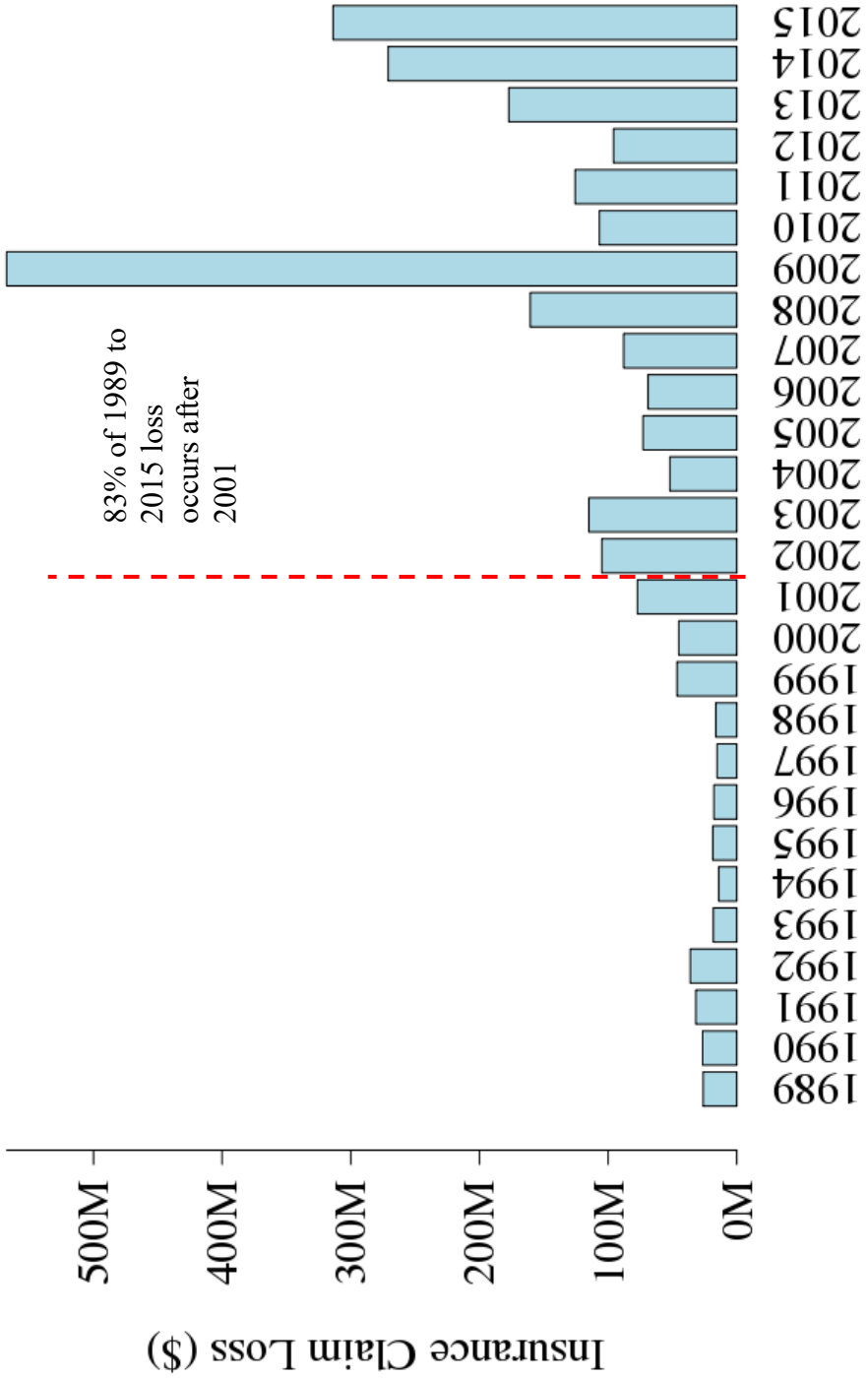


Figure 1.2. Total insurance loss by year for the three-state region of Washington, Oregon, and Idaho, from 1989 to 2015. Losses from 2001 to 2015 make up 83% of total losses from 1989 to 2015.

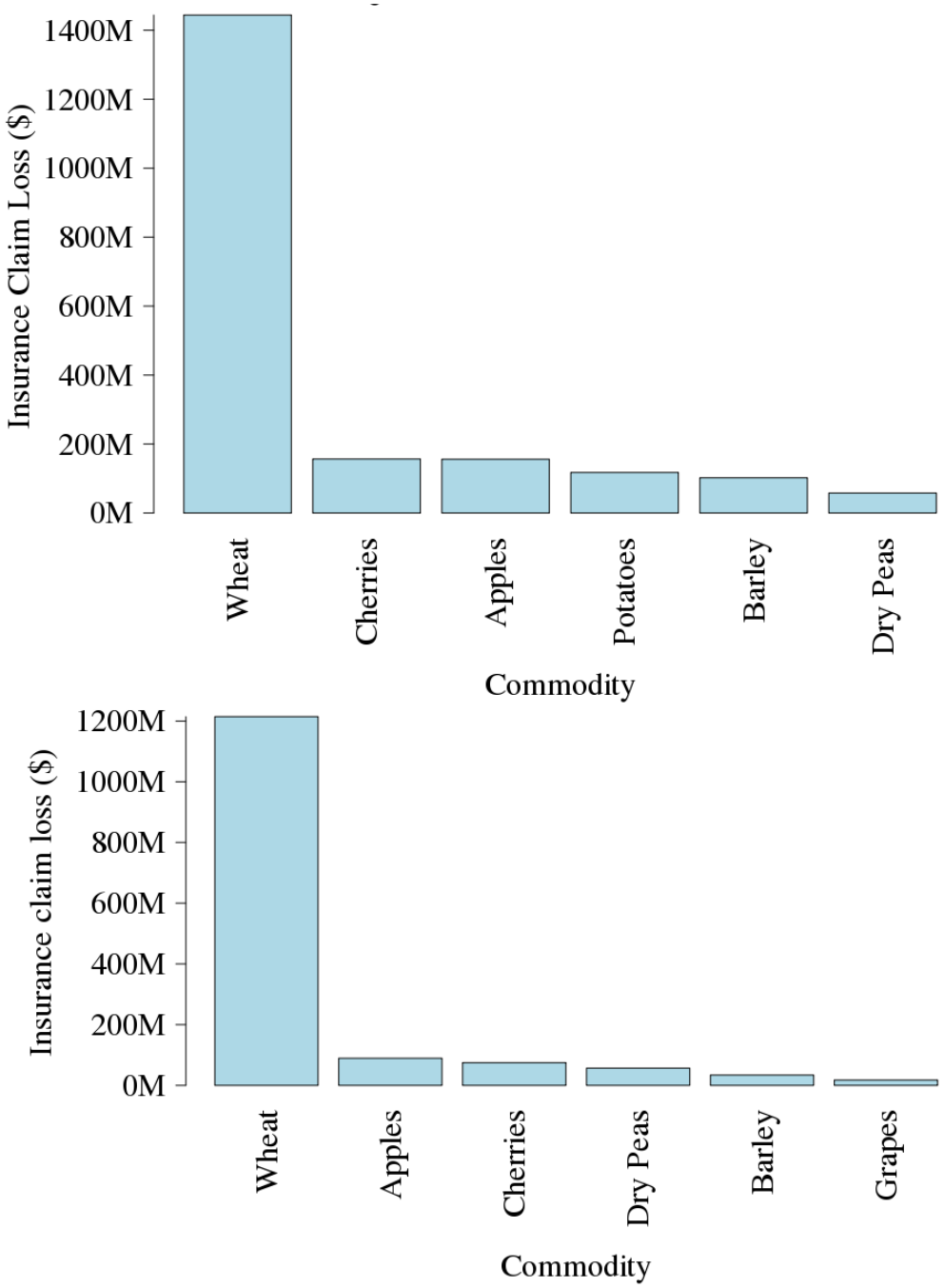


Figure 1.3. Total insurance loss (2001 to 2015) for the top six commodities for the PNW (top), and the iPNW (bottom).

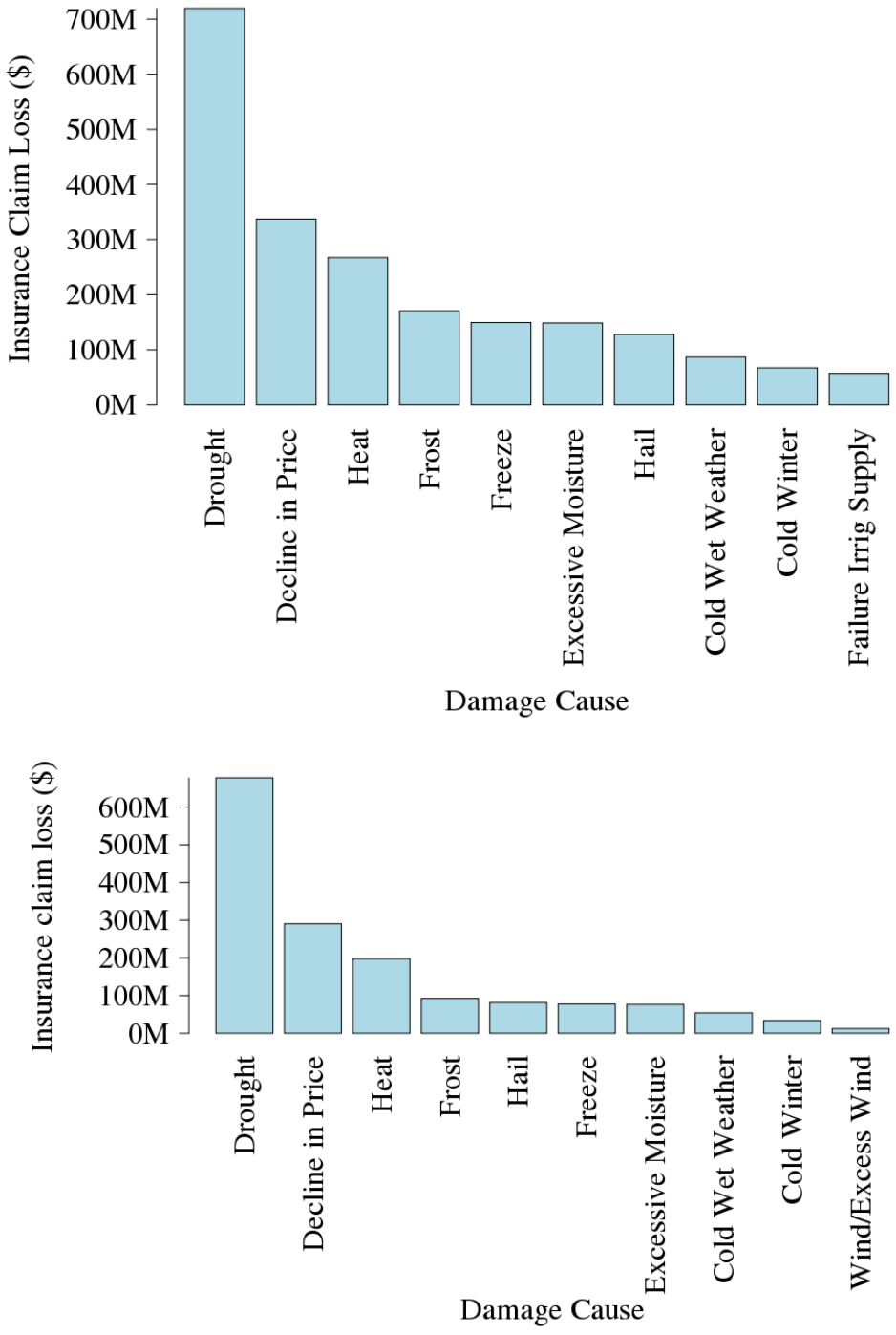


Figure 1.4. Total insurance loss (2001 to 2015) for the top ten damage causes for the PNW (top), and the iPNW (bottom).

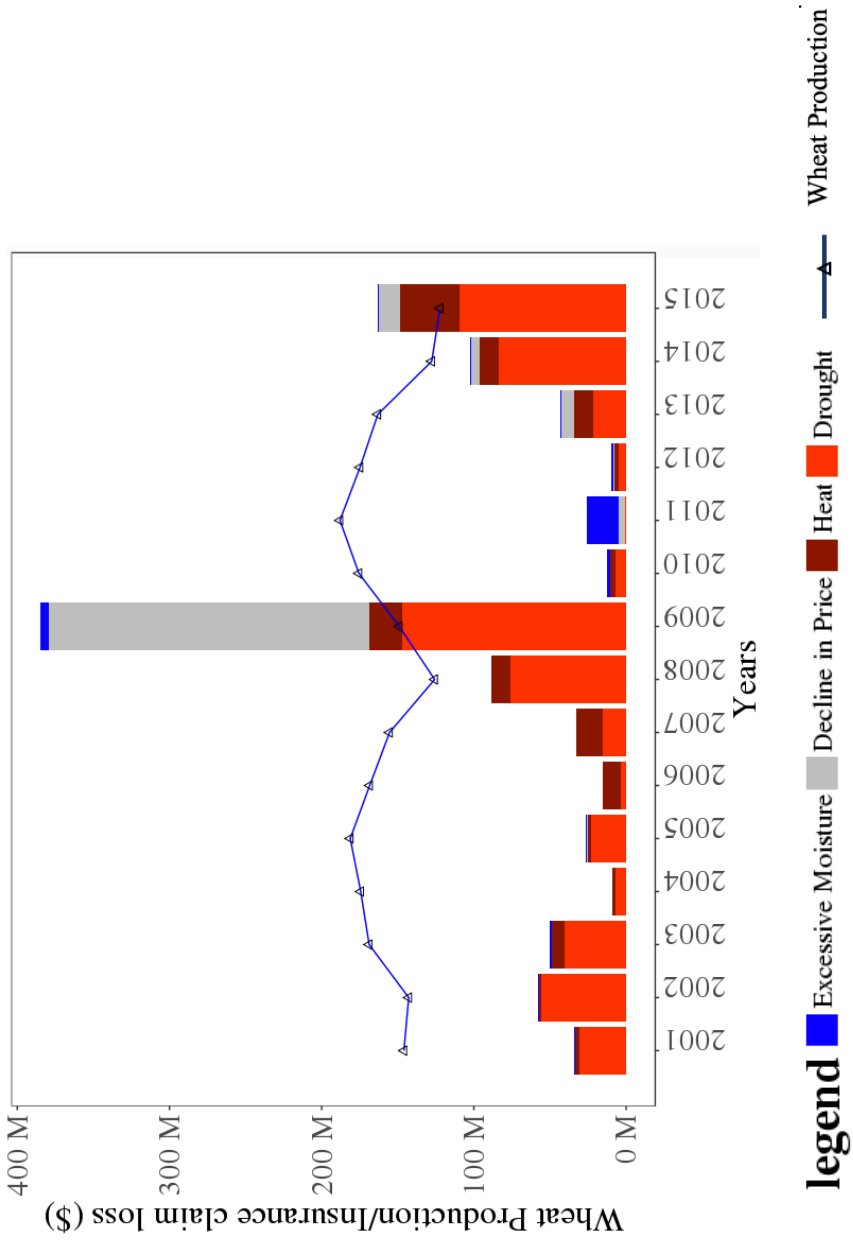


Figure 1.5. Wheat commodity loss ($) for the iPNW, showing the top damage causes (excessive moisture, drought, heat, and decline in price) compared to wheat production (National Agricultural Statstics Service [NASS]) for each year, from 2001 to 2015.

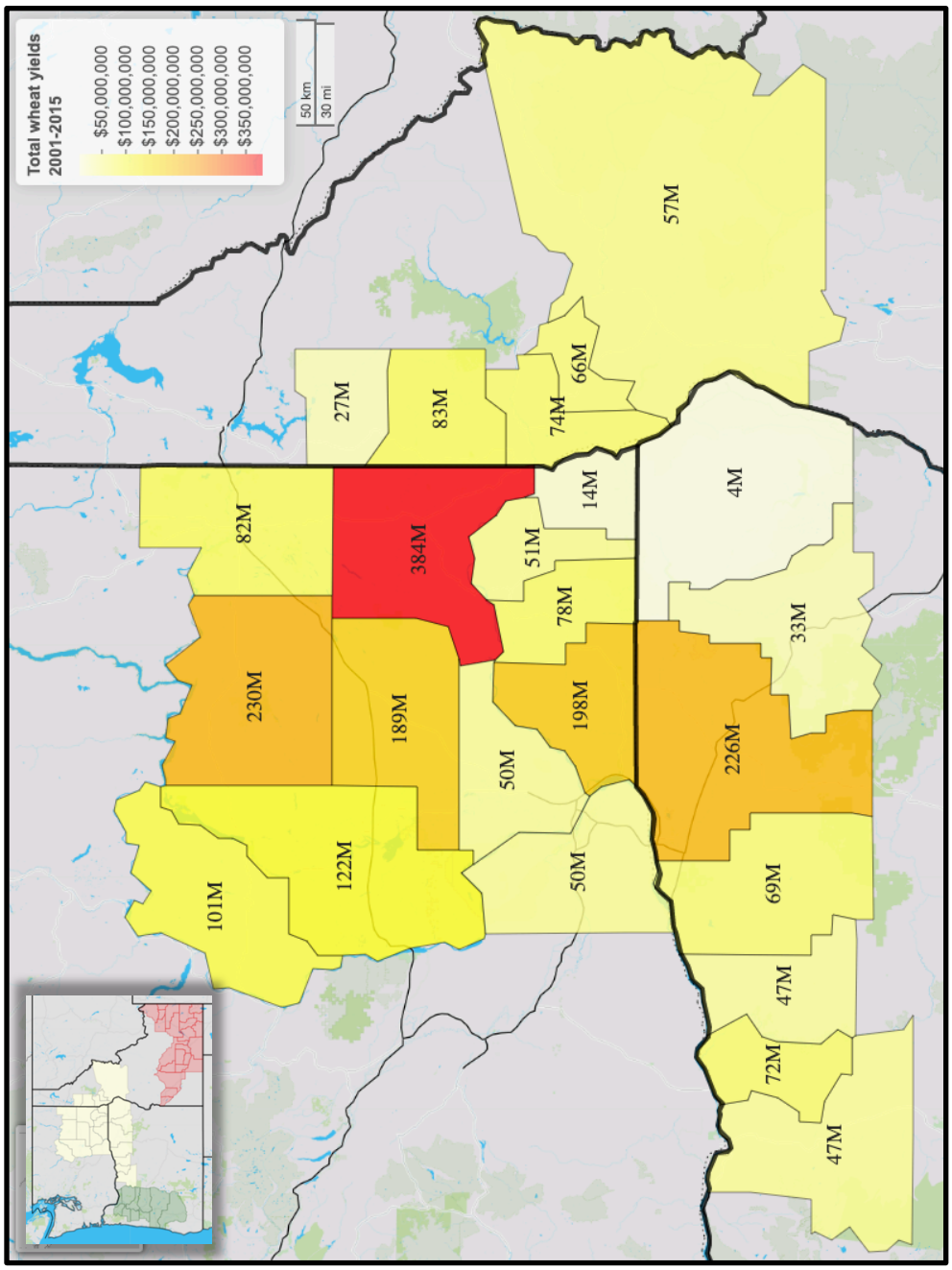


Figure 1.6. Map of total yields ($) per county for wheat, from 2001 to 2015 (NASS). Values on map listed in millions of dollars.

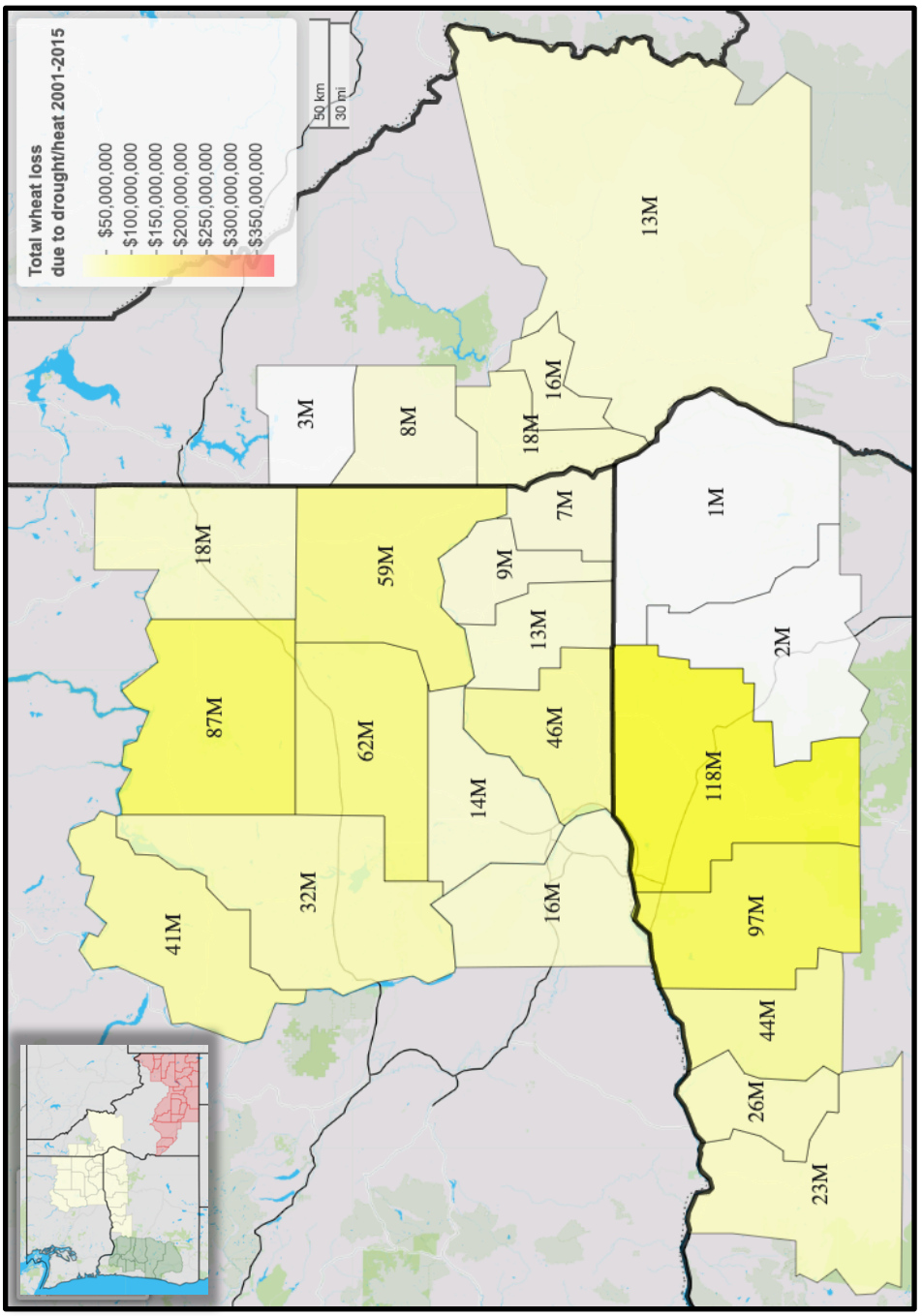


Figure 1.7. Map of total insurance loss ($) for wheat, due to drought and heat, from 2001 to 2015. Values on map listed in millions of dollars.

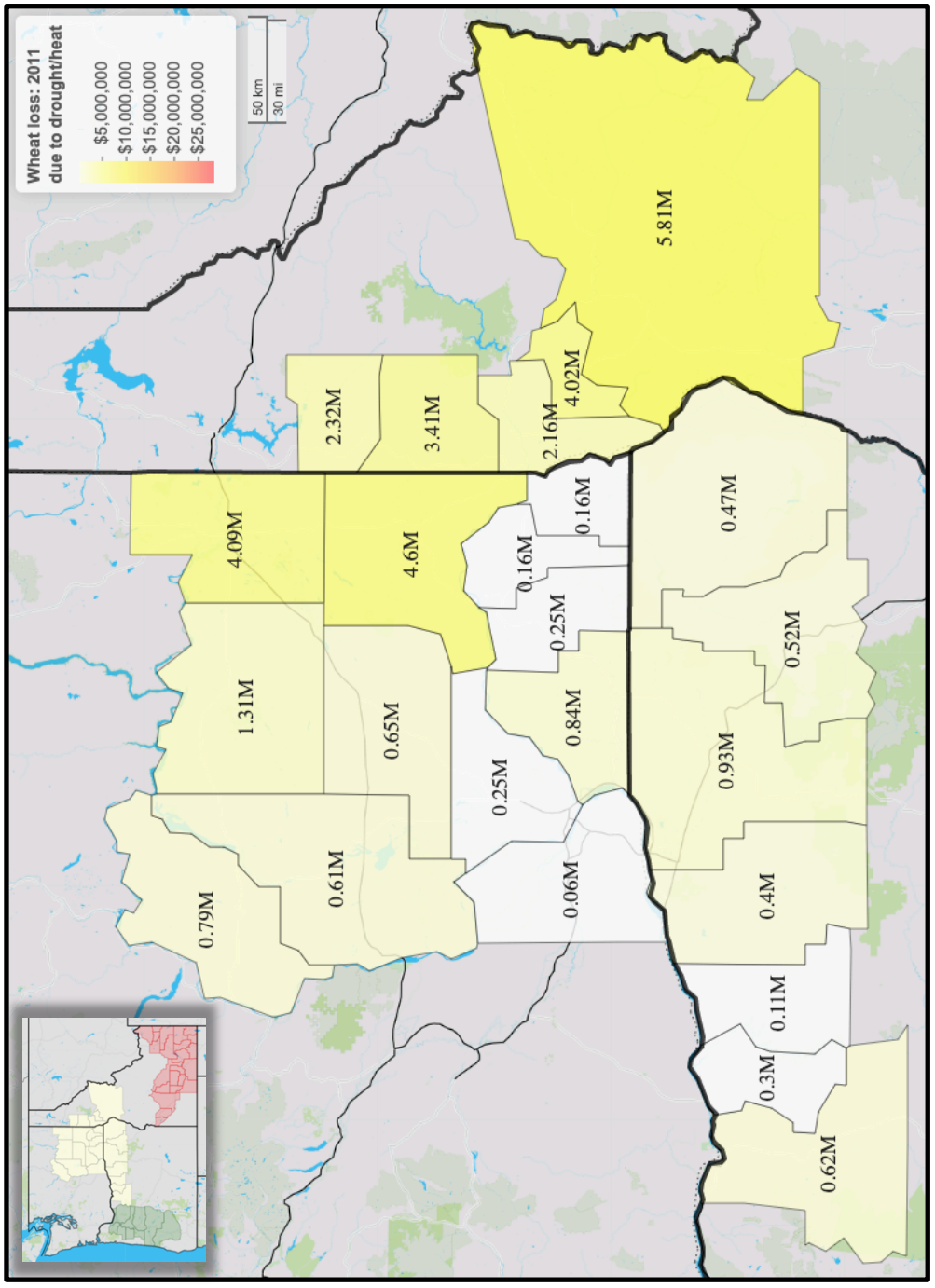


Figure 1.8. Map of wheat insurance loss ($) due to heat and drought, for 2011.

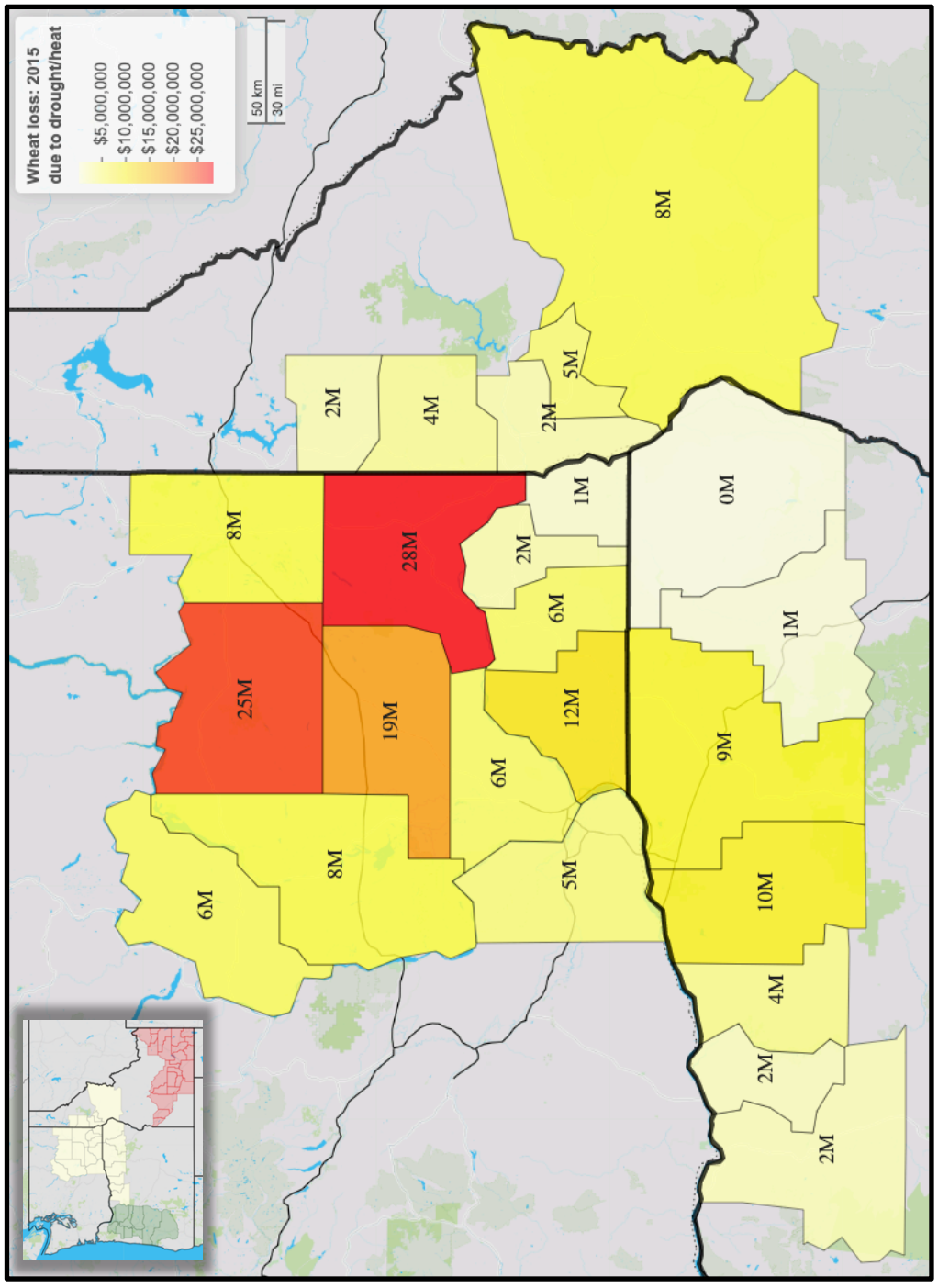


Figure 1.9. Map of wheat insurance loss ($) due to heat and drought, for 2015.

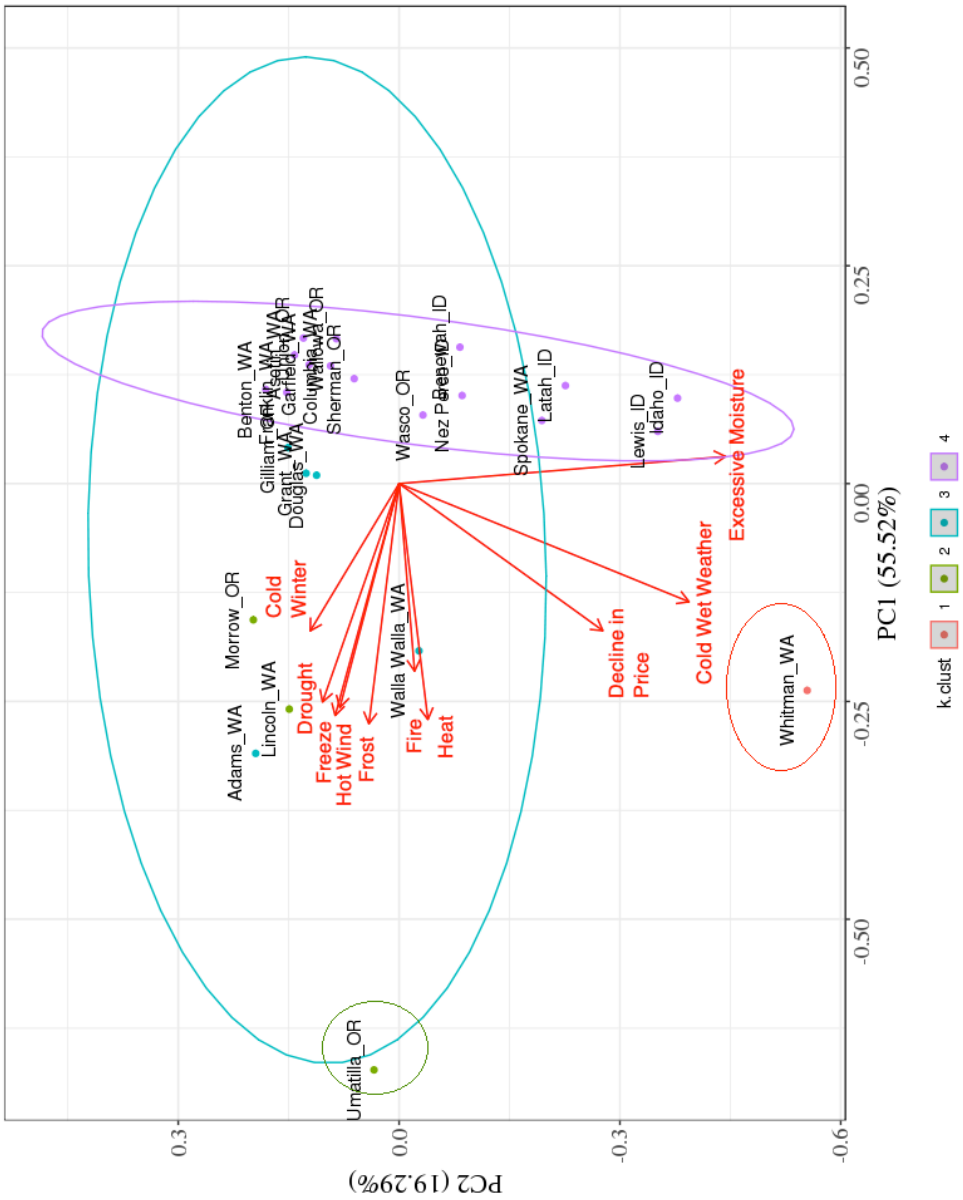


Figure 1.10. Principal components analysis (PCA) showing top damage cause factor loadings for iPNW wheat insurance loss, from 2001 to 2015, with counties as the independent variable. The top two principal components account for approximately 75% of the overall variance. Clustering was constructed using a kmeans technique.

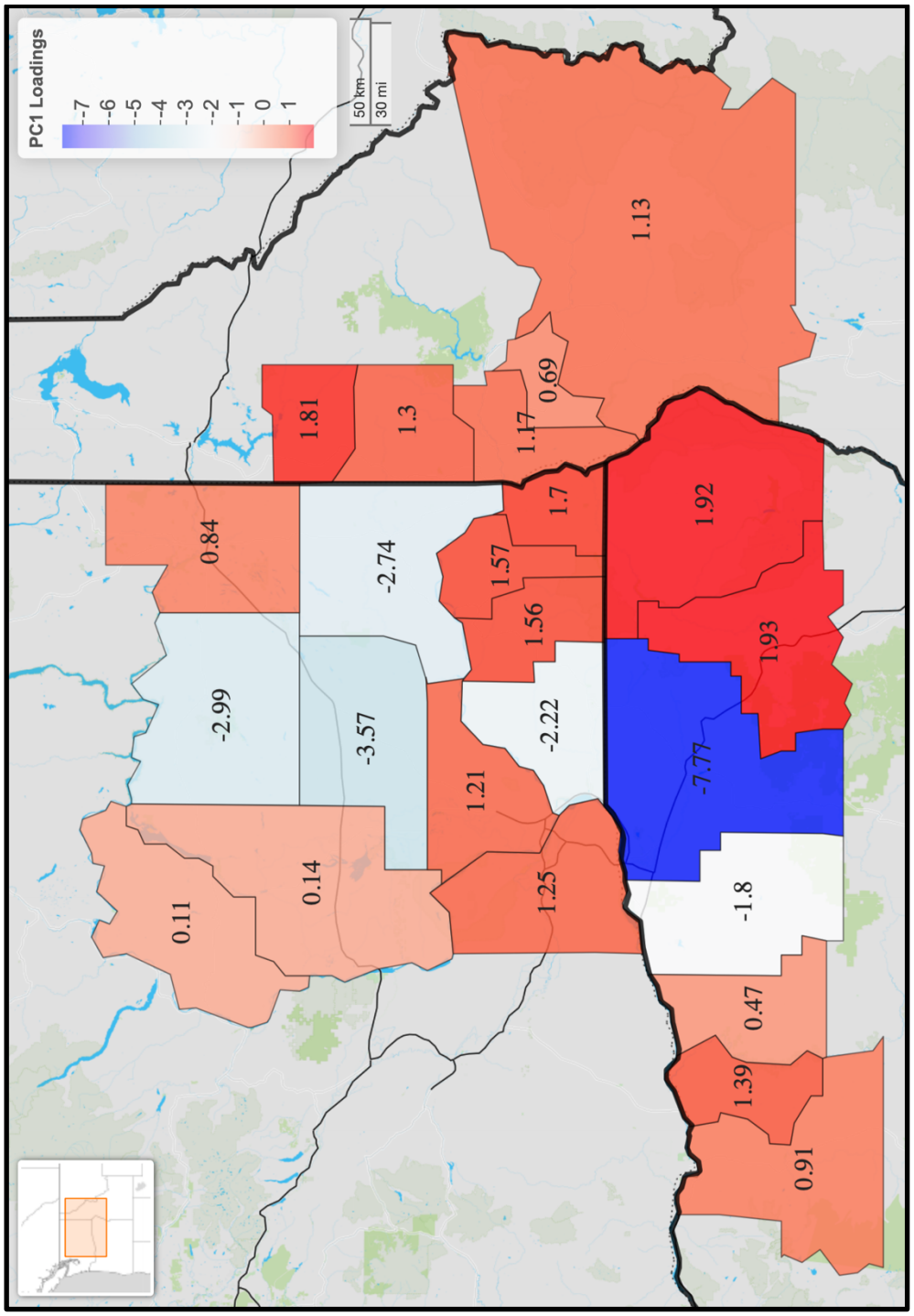


Figure 1.11. Map of PC1 loadings for wheat by county, based on damage cause factors, 2001 to 2015.

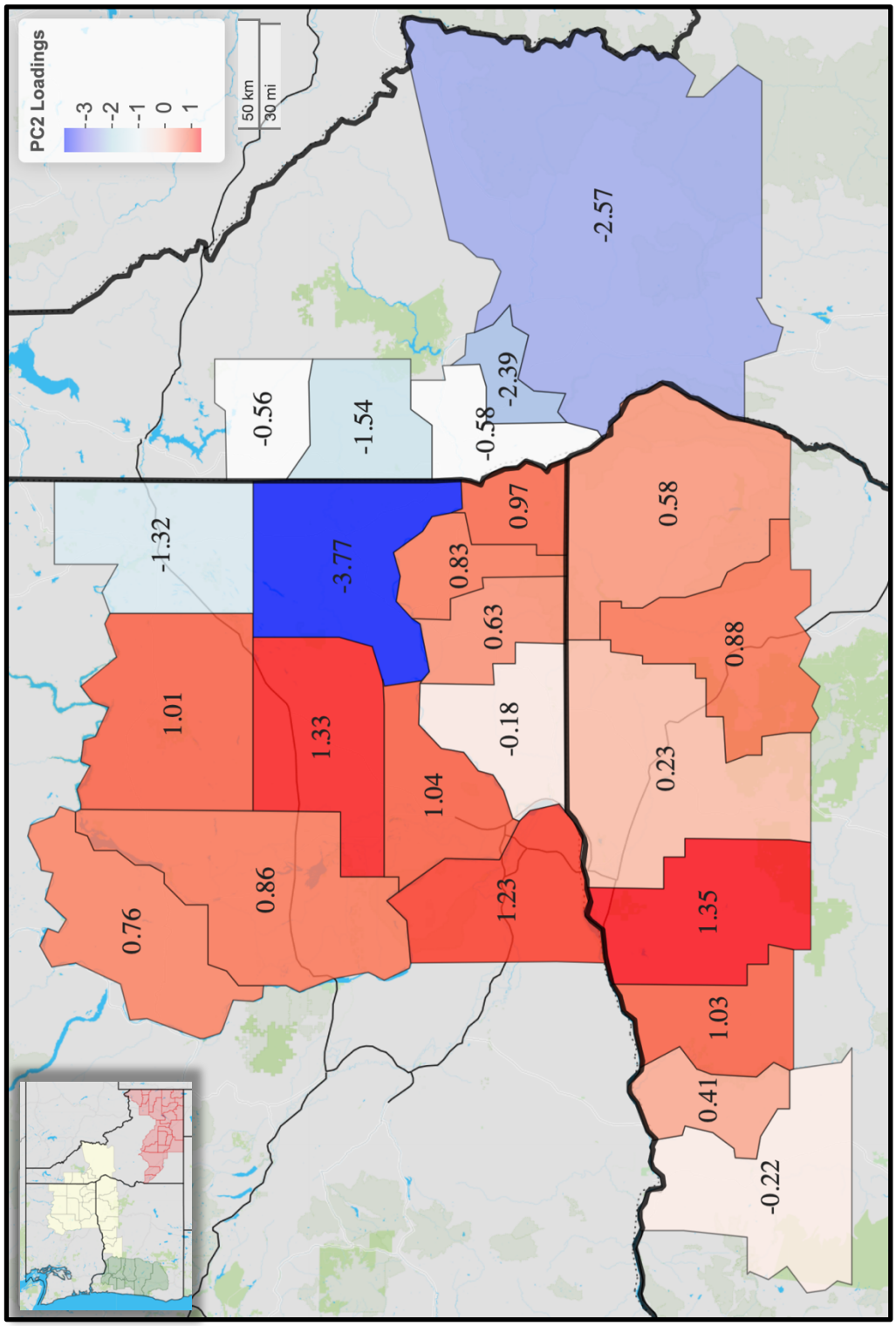


Figure 1.12. Map of PC2 loadings for wheat by county, based on damage cause factors, 2001 to 2015.

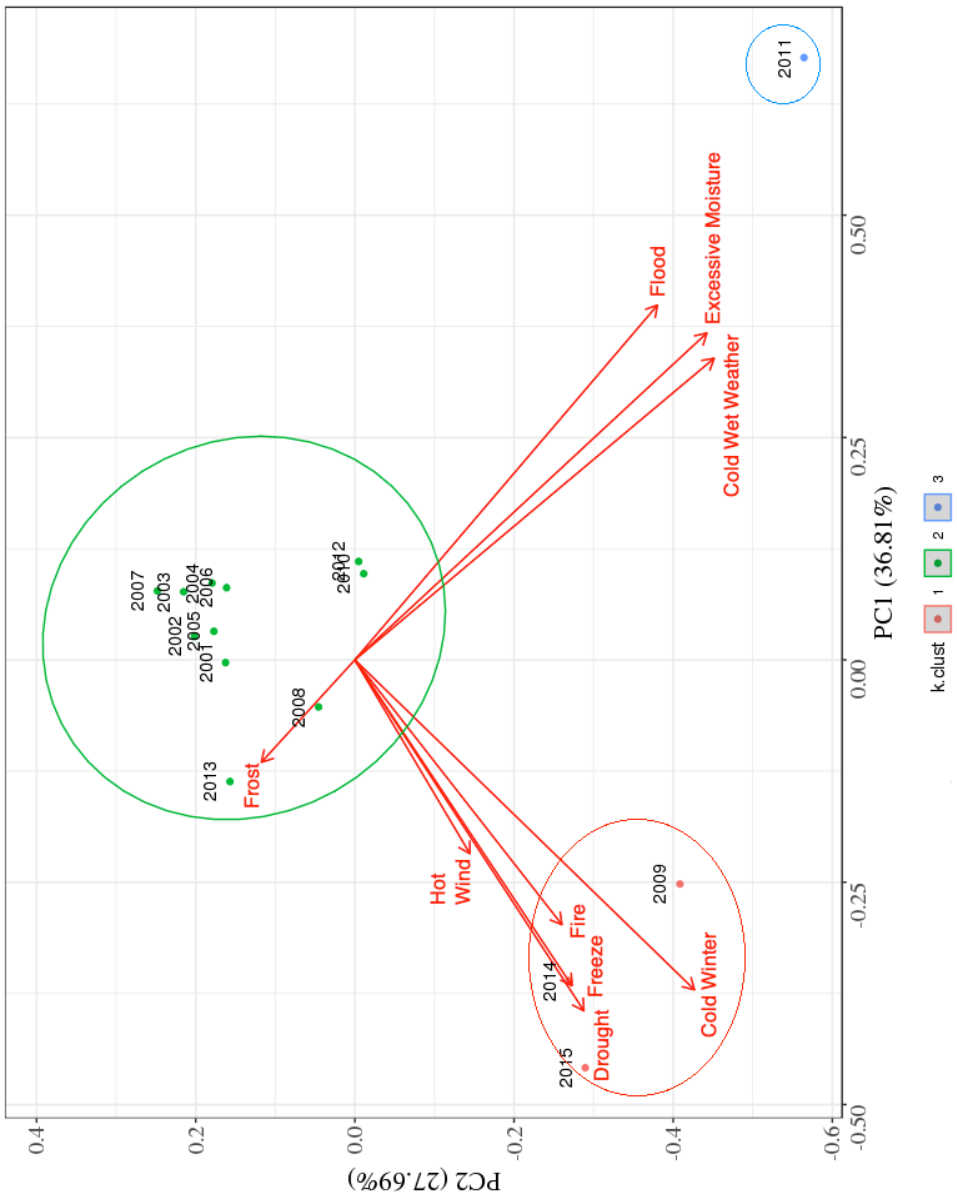


Figure 1.13. Principal components analysis (PCA) showing top damage cause factor loadings for iPNW wheat insurance loss, from 2001 to 2015, with year as the independent variable. The top two principal components account for approximately 64% of the overall variance. Clustering was constructed using a kmeans technique.

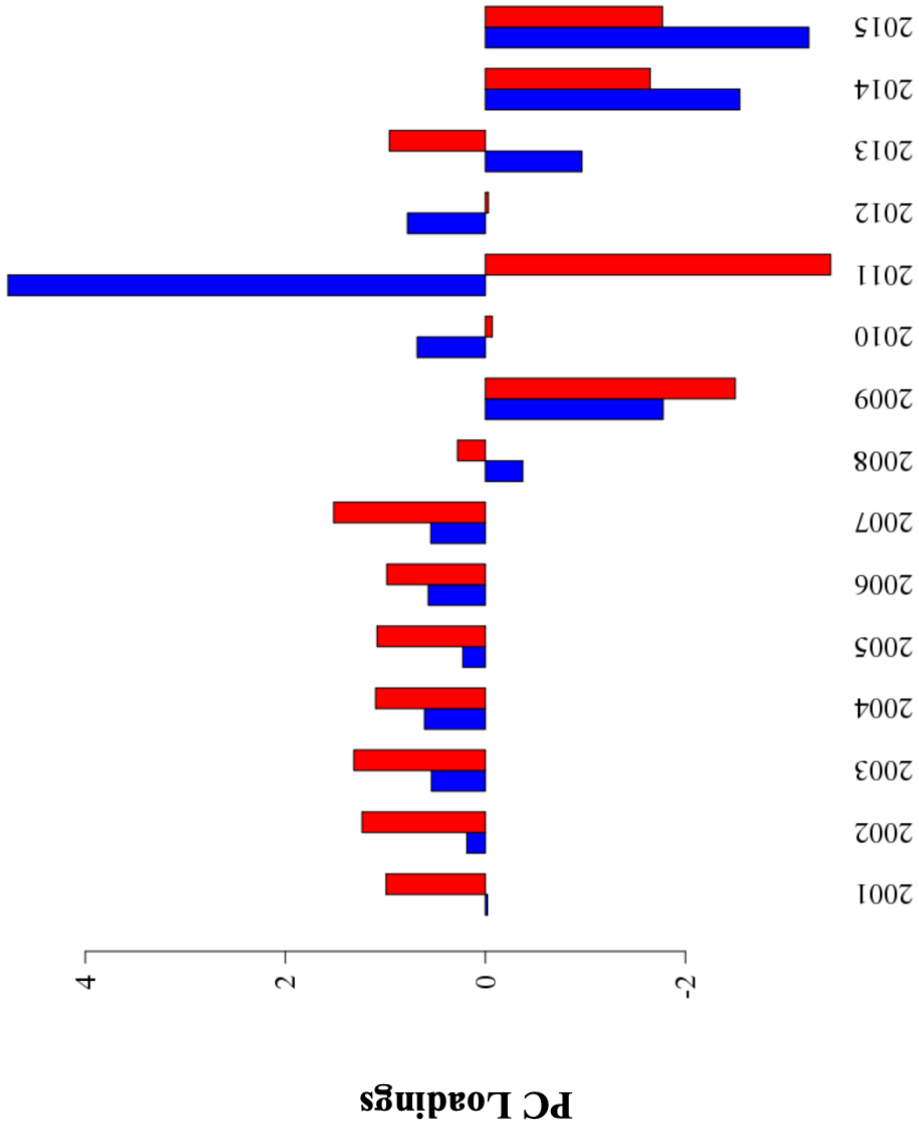


Figure 1.14. IPNW insurance loss PC loadings by year.

CHAPTER 2: REGRESSION BASED RANDOM FOREST MODELING OF INLAND PACIFIC NORTHWEST DROUGHT-RELATED WHEAT INSURANCE LOSS USING TIME LAGGED CLIMATE CORRELATION MATRIX ASSOCIATION

2.1. Introduction

Climate change significantly adds to the challenges facing agriculture, such as ensuring food security and preserving the economic prosperity of a growing global population (Deschênes & Greenstone, 2018; Schlenker & Roberts, 2009; Fan et al., 2016; Epstein, 2005; Rosenzweig et al., 2001; Manandhar et al., 2014). Of noted interest are the varied effects of climate on weather-related phenomena (e.g. drought, heat waves, flood, hurricanes, extreme precipitation) (Trenberth, 2000). Drought in particular affects a number of factors associated with cropping systems, including excessive temperatures, water availability, long term soil moisture drawdown, and levels of evapotranspiration (Lobell & Costa-Robert, 2011; Rosenzweig & Parry, 1994). This analysis focuses on the highly productive inland Pacific Northwest (iPNW) agricultural region of the United States (figure 2.1), which relies heavily upon dryland farming for cereal production (Karimi et al., 2017; Yorgey & Kruger, 2017), and is considerably impacted by water availability. Our research objectives were to model how climatic effects for the iPNW are related to agricultural insurance loss, with a particular focus on drought related claims in wheat. Two questions are addressed: What climate variables and temporal windows best relate to drought claims for wheat, and how do these relationships vary across the iPNW? Additionally, based on these optimum relationships, what climate variables have a greater influence on wheat insurance loss due to drought for the region, and can we utilize this framework for the prediction of insurance loss?

2.2. Motivation

From a regional perspective, the iPNW produces around 17% of the U.S. wheat harvest (Karimi et al., 2017; Roesch-McNally, 2018). Given that cereal production is directly linked with variations in precipitation and temperature across much of the U.S. and globe (Hatfield & Dold, 2018), much of the iPNW wheat yields are significantly correlated with variability in plant available water during the growing season (Yorgey & Kruger, 2017; Chi et al., 2017). For example, 2015 iPNW wheat cropping outputs were negatively impacted by drought and extreme temperatures, evidenced in reduced crop yield outputs, increased agricultural insurance loss claim totals, and overall reduced wheat quality resulting in approximately $200 million of insurance loss in Washington state alone (Seamon et al., 2019a, Sandison, 2017; Howitt et al., 2015).

Mishra and Singh (2010) define agricultural drought as "a period with anomalously low soil moisture that substantially limits crop production are extended over multiple years". During such conditions, the ability for landscapes to recharge water storage becomes more difficult, effecting vegetation cycles, soil erodibility, as well as agricultural practices that can severely impact farmer outputs (Lobell & Burke, 2010). While iPNW drought conditions have historically been driven by a lack of precipitation, warming temperatures have played an increasing role on future regional water availability and drought (Abatzoglou et al.., 2014; Marlier et al., 2017; Mankin & Diffenbaugh, 2015). In addition, composite or index variables can also be useful in quantifying drought conditions, such as the Palmer Drought Severity Index (PDSI), which attempts to standardize drought impacts across differing climates and is dependent upon the available water holding capacity of soils (Palmer, 1965). While agricultural practices such as crop rotations, planting cycles, genetic selection, and

pest/fertilizer/water management have improved agricultural system efficiencies over the past 30 years, these improvements are partially offset by more adverse growing conditions in some regions given the impacts of climate change (Hatfield et al., 2014).

In terms of mitigating long term and seasonal variability with regards to climatic impacts, crop insurance is a key mechanism that is used to reduce such risk (Miranda et al., 1997; Seamon et al., 2019a). Of particular note are long-term climatic variability impacts on short-term extreme weather events, as well as shifts in seasonal/subseasonal weather outcomes that are exacerbated by a changing climate over an extended period of time (USGCRP, 2017). In the PNW alone, from 2001 to 2015, over 35,000 insurance claims were filed, of which 20,600 were for wheat (Seamon et al., 2019a). For the iPNW in particular, drought and heat insurance claims for all commodities resulted in approximately $760 million in insurance losses from 2001 to 2015, which account for approximately 55% of all losses for this time period (Figure 2.2). Of particular interest in terms of climate, agriculture, and crop insurance, are the individualized phenological cycles per cropping type and their seasonal relationships, particularly in regard to insurance loss outcomes. Due to the diverse nature of a particular crop's growth cycle and its varying sensitivities to climatic effects, the seasonal timing of long-term climatic changes can be just as important as the extreme nature of a particular weather event (Barlow et al., 2015; Hatfield et al., 2014, Asseng et al., 2013). Long term climatic changes on cropping systems, given non-linear relationships, may result in alterations in yields or losses after surpassing particular thresholds (Hatfield et al., 2014). Cropping systems have differing responses to temperature changes throughout the phenological life cycle, with most species typically requiring higher temperatures (presuming adequate water availability) for optimum vegetative growth rather than for reproductive

development (Hatfield & Prueger, 2015). As a result of these life cycle temperature needs, extreme climatic events will have differing effects on differing crops based on their phenological stage. These extreme event outcomes may result in shorter life cycles, lower plant production, reduced reproductive periods, decreased pollen viability, and lower yields. As such, these impacts may result in increasing the overall risk of financial loss, and therefore, increase the potential for higher insurance losses, as well as the frequency of claims.

Grain-based cropping systems are particularly impacted by potential increased temperatures. Considerable research has examined the range of temperature impacts on grain yields (Sacks & Kucharik, 2011) indicating that progressive temperature increases may initially result in increased yields, with an accelerating decrease over time, given an inverse temperature/precipitation relationship. Stockle et al. (2018) note that while increased temperatures will likely decrease wheat yields in the region, the effects of CO_2 fertilization may modestly offset these yield reductions over time. In contrast, Schlenker and Roberts (2009) suggest that yields for alternative forms of cropping systems, such as soybeans, corn, and cotton, would slightly increase with initial temperature increases up to 32 degrees Celsius, and then sharply decrease as temperatures rise above that threshold. To make matters more complex, Rezaei et al. (2018) as well as Asseng et al. (2013) indicate that unique cultivars within a species may have varying phenological cycles, suggesting that any agricultural climate impacts assessment should include a variety of sub-species for proper threshold analysis. When examined in total, climatic relationships to agriculture are extremely variable, with changing outcomes due to cropping system, regionalization, farming practices, and genetic diversity. This complexity is encapsulated in agricultural insurance

loss management, in order to effectively hedge agricultural risk, associated variability and complexity, and incorporated into a time-adjusted financial premium/payout process. Under this premise, evaluating insurance losses in relationship to subseasonal climatic impacts provides a streamlined approach to assess patterns and predictability, without delving into the underlying crop processes and their biophysical effects due to a changing climate.

Given this relationship between climate and agriculture, our research focus was to determine which climate variables and temporal windows best relate to drought claims for wheat, and what spatial variability exists with regards to this climate/agriculture association. Armed with this optimized information, we then developed a predictive model for estimating agricultural insurance loss based on climatic influences.

2.3. Data and Methods

The study area for this analysis was a 24-county region of the inland Pacific Northwest (iPNW) of the United States (Figure 2.1). As one of the key agricultural production regions in the U.S, it supports a variety of cropping systems and management practices, which are dominated by dryland wheat farming (Chi et al., 2017). The region has cool to cold, wet winters and warm to hot, dry summers, with considerable interannual variability based on regionalization across the PNW three state area (Yorgey & Kruger, 2017; Abatzoglou et al., 2014). Agricultural production for the area is typically limited by water rather than by growing season length (Stockle et al., 2018), with annual precipitation increases from west to east, ranging from 200mm to over 600mm (Schillinger et al., 2010; Chi et al., 2017). Interannual climate variability in the region can lead to considerable variation in available

moisture, temperature, and evaporative demand, all of which can be influential on long term agricultural systems production.

Two datasets were organized for this analysis: 1) the USDA agricultural crop insurance claim archive from 2001 to 2015 (http://rma.usda.gov). When U.S. farmers experience economic losses for particular agricultural commodities, they typically file a crop insurance claim for that loss, as part of the Federal Crop Insurance Corporation's (FCIC) program, which underwrites agricultural insurance policies in conjunction with private insurance organizations. The filing of a crop insurance claim is the result of complex decision making process, where farmers may incorporate multiple factors, spanning biophysical, climatic, economic, and socio-demographic disciplines. Over time, these insurance claim records are systematically provided to the U.S. Department of Agriculture (USDA), who administers the program via the USDA Risk Management Agency (http://usda.rma.gov) and makes these data available as a public archive. Individual records are spatially coarsened to provide anonymity, by removing address information and providing the data only at a monthly, county-level temporal/spatial scale. Each individual insurance claim indicates the commodity insured, the year and month of filing, the cause of damage, as well as the total amount claimed (in dollars). The USDA agricultural insurance dataset contains individual claims (with the dollar amount of the claim loss) based on four key factors: time (the year and month of the claim), the state and county the claim was filed in, the commodity type, the acreage of the claim, and the specific, singular cause of the claim. For the purposes of this analysis, individual agricultural insurance claims for wheat due to drought were aggregated to the county level. We only included claims during January to September given our focus on winter wheat phenological cycles, and hereafter focused on annual claims for each county.

The second dataset used was daily gridded climate data at 1/24th degree (4km) spatial resolution for three climate variables that are most associated with water availability (maximum temperature, precipitation, and potential evapotranspiration) from Abatzoglou (2013). Daily values were summarized by county and to a monthly time scale, with totals aggregated for precipitation and potential evapotranspiration, and average values for maximum temperature and PDSI.

In order to explore the associations of climate to insurance loss, we constructed a set of time lagged climate variable design matrices, to search for the optimal seasonal temporal relationship between a climate variable (i.e. maximum temperature, precipitation, evapotranspiration, and PDSI) and a county's seasonal wheat insurance loss due to drought. We used this optimization approach to identify the most influential county level time periods per season that were best correlated to wheat/drought insurance claims. Through these steps we evaluated a variety of machine learning algorithms individually, as well as part of an ensemble process, to initially evaluate predictive performance (singular regression decision tree, support vector machines, random forest, artificial neural networks), using a common out of sample/in sample structure (10-fold cross validation). Based on the RMSE/R^2 results of the aforementioned algorithm outputs, we chose a regression-based random forest approach to predict insurance loss annually over a time span of 2001 to 2015. A particular advantage of random forest techniques is the ability to evaluate the feature importance of climatic predictors (Liaw & Wiener, 2002), as well as the minimization of model overfitting (Brieman, 2002).

Our initial step was to perform a general comparison of county-level wheat/drought related claim acreage to water availability variables (precipitation and potential evapotranspiration), as well as aridity (precipitation / potential evapotranspiration), to examine general trends over time (2001 to 2015). The purpose of this initial analysis was to verify that expected patterns of wheat/drought insurance loss acreage vs. water availability were seen (e.g. counties with relatively lower precipitation totals had overall increased wheat/drought insurance loss acreage). These annual, county-specific ratios were calculated by taking the total number of acres of wheat/drought insurance loss and dividing by the total number of acres for all other wheat damage cause losses. As the ratio approaches 1, we identify county/year combinations where drought was the dominant factor in terms of total farmland acreage attributed to insurance loss.

Secondly, we aggregated and associated climate data to wheat/drought crop loss seasonal totals by county, analyzing all outputs by using a time-lagged association by searching for the highest correlations between all combinations of monthly climate values (scaled and mean centered) and wheat insurance loss claims due to drought. We transformed annual loss amounts by a cube root function which resulted in a normal distribution of annual values. Following previous efforts to elucidate the optimum correlation (e.g., Du et al., 2013), we examined these climate/insurance loss correlations for each county, based on each climate variable, using a 12 x 9 design matrix that considered different temporal periods (12 months for all climate data, 9 months for only winter wheat claims from January to September). Our goal was to evaluate which time-lagged windows were best correlated with overall wheat/drought insurance loss for a county, across the 2001 to 2015 time period. Using the results of this optimization, we then combined results of our county-specific time lagged

correlation data across the 24-county study area, examining coefficients of correlation (R). The output of our design matrix methodology resulted in an optimized county level climate/insurance loss dataset (wheat claims due to drought), for each year, for the entire study area. This process created a new dataset of dependent and optimized independent (maximum temperature, precipitation, potential evapotranspiration, PDSI) variables, which were used in the final step of our analysis: a regression-based random forest/decision tree analysis.

Regression decision trees are a method of constructing a set of decision rules on a predictor variable (Breiman et al., 1984; Verbyla, 1987; Clark & Pregibon, 1992) that is continuous (versus categorical). These rules are constructed by recursively partitioning the data into successively smaller groups with binary splits based on a single predictor variable, with the goal of encapsulating the training data in the smallest possible tree (Prasad et al., 2006). The rationale for minimizing the tree size is the premise that the simplest explanation for a set of phenomena is preferable over other explanations (Behrens, 2000). Regression decision tree development initially grows maximal trees and then uses techniques such as cross validation/rotation estimation to prune the overfitted tree to an optimal size (Therneau & Atkinson, 1997). Regression decision trees have several advantages over traditional statistical methods, including: uncovering conditional structure in data with hierarchical variables, assessing non-linearity given that no a priori knowledge is assumed, as well as providing insight into spatial tendencies given its ability to map predictors with the greatest influence on a distribution (Prasad et al., 2006). Computationally, decision tree processing is typically based on either the ID3 or C3.4 algorithms (Quinlan, 1986), which are used to

construct a decision tree for regression by replacing information gain with standard deviation reduction.

Random forest, or ensemble decision trees, are a combination of many decision tree predictors, where each tree depends on the values of a random vector, sampled independently and with the same distribution for all trees in the forest (Brieman, 2001). Random forest modeling reduces the potential for overfitting through the use of bootstrap aggregation, averaging across many trees, and provides a level of feature importance for assessing predictor power. As part of our random forest construction, we utilized 10-fold cross validation, a model validation technique used to assess the generalizability of the model. Model construction for this analysis utilized the recursive partitioning and regression trees package (rpart) within R (Therneau & Atkinson 2011; Breiman et al., 1984). To tune the model, we performed hyperparameter optimization as part of our cross validation, which iterates over the model to find the best set of parameters for optimization and minimizes the loss/error function (in regression, typically the mean squared error (MSE) or the root mean squared error (RMSE). We then examined this error using learning curve theory, which compares the performance of a model's training and testing data over a varying number of training instances, with the expectation that model performance improves with an increase in observations. By separating the training and testing datasets and plotting them individually as the model is run repeatedly, while increasing observations, we can get a sense of how well the model will generalize with new data (i.e. changing climate variables or a change in commodity pricing). Additionally, learning curve analysis allows for the evaluation of the balance between bias and variance, by examining how the curves converge or persist in separation (Perlich et al., 2004).

2.4. Results

The results of our initial county level wheat/drought acreage ratios (2001 to 2015), as compared to precipitation, potential evapotranspiration, and aridity are shown in Figure 2.3. Each observation represents the mean annual totals of each individual variable, for each county, from 2001 to 2015. We see expected climate/insurance loss relationships, with county level precipitation totals, as well as aridity, inversely proportional to increased drought acreage ratios. Similarly, potential evapotranspiration totals increased as ratios increased. These general patterns supported moving forward to explore more specific spatiotemporal relationships of climate to insurance loss using our climate-lagged correlation framework.

2.4.1. Climate vs insurance loss time-lagged relationships results

Using our time-lagged matrix approach, we identified the optimum correlation between annual wheat/drought insurance loss and each variables' climate window, for each county (Figure 2.4 shows an example of this climate lagged matrix design for Whitman county, WA). Using these time-lagged matrix correlation outputs, we mapped the spatial and temporal variations for each climatic variable, at a county scale. This mapping allowed us to identify notable patterns that align to physiographic features, such as elevation. In addition, we plotted these optimum temporal windows for each county, and for each variable. Finally, we constructed bivariate plots that compared each county/year combination with the related wheat/drought insurance loss, with wheat price used to vary the observation point size. These three analysis views are organized for each of the predictor climate variables.

Potential Evapotranspiration (PET). An overview analysis of PET from 2001 to 2015 for the study area indicates a climate gradient from east to west, with the annual maximum occurring in July (Figure 2.5). When examining optimum correlations of PET to cube root transformed wheat/drought insurance loss (Figure 2.6), the counties of Walla Walla, Columbia, and Whitman (Washington) had the highest values, along with Latah and Benewah counties (Idaho). Southernmost counties (Oregon) tended to have the lower correlations with earlier seasonal months (Feb/Mar/Apr). In terms of the optimum time windows for PET (Figure 2.7), we see a shift toward later time windows from southwest, through Washington into Oregon. The optimal time window for Idaho counties tended to be earlier in the season (March/April/May), while that window is slightly later for Washington counties (April/May/June/July/Aug), and for Oregon counties that time window is later still (July/August/September). When we analyze the correlation of all optimal county/year climate combinations with insurance loss (Figure 2.8), we see moderate positive correlations (R = .49), with similar regionalized groupings between Idaho, Oregon, and Washington counties: Idaho counties tended to have higher precipitation with lower wheat/drought insurance loss, with Washington having the highest insurance loss totals with the highest levels of PET.

Precipitation. Regional cycles of precipitation, averaged for the period from 2001 to 2015, show increasing values from northwest to southeast. In all counties, the wettest months are November through January, but spring (March-April-May) is wet too (Figure 2.9). Negative correlations of precipitation (Figure 2.10) with insurance loss were highest in the southernmost study area counties, which tended to border the Snake and Columbia rivers (Wasco, Sherman, and Union counties, OR, and Columbia and Whitman counties, WA),

ranging from -.61 to -.81. Counties further to the northwest (Franklin, Douglas, Lincoln, Spokane counties, WA) tended to have lower correlations, with April/May/June precipitation more highly correlated with drought loss for most counties (Figure 2.11). However, the counties furthest west (Benton, Grant, Douglas), which are fruit dominant regions, had differing time windows for optimal insurance loss associations. When combined in an overall bivariate plot (Figure 2.12), we saw modest inverse correlations with precipitation (R = -.44), with increased precipitation in counties with overall lower wheat/drought insurance claim totals. Similar to PET, we see regionalized groupings for study area counties by state, with Idaho counties having the highest precipitation totals and the lowest overall wheat/drought insurance loss.

Maximum Temperature. Temperature tends to increase from southeast to northwest, which, as expected, is opposite to the pattern for precipitation. July was the peak month for high temperatures for every county in the iPNW (Figure 2.13). Maximum temperature was most highly correlated with drought (Figure 2.14) in southeastern Washington counties, including Walla Walla, Whitman, Columbia, Garfield, and Spokane counties, WA, as well as Latah county, ID, ranging from .59 to .72. Temporally, most counties had an optimum relationship that fell between April and July, with Oregon counties and several western WA counties shifting that time window slightly later (Figure 2.15). The two counties with the highest correlations (Columbia and Garfield counties, WA), both had optimum time windows that included April. When all county/year combinations are examined in a bivariate plot (Figure 2.16), we see R = .47, with a positive correlational distribution, indicating higher loss with increased temperatures.

Palmer Drought Severity Index (PDSI). PDSI spatial gradients had an increasing trend from southeast to northwest (Figure 2.17). PDSI optimum correlations with wheat/insurance loss appeared to have a spatial gradient that similarly increased to the northwest (Figure 2.18), with Douglas, Benton, and Grant counties, WA, as well as Wasco county, OR, having the highest correlation coefficients (R = .81). July/August/September was the most common window for optimum climate/insurance loss relationships (Figure 2.19), yet those counties with the highest correlations had window ranges that were much longer. As with precipitation, PDSI is negatively correlated with wheat/drought loss (Figure 2.20).

2.4.2. Regression based random forest modeling results

The foregoing correlations demonstrate significant lagged effects of late spring climate variables on drought loss. Building on this understanding, we uncover nonlinear effects using random forest modeling. The overall results of our 10-fold cross validated random forest model using insurance dollar loss totals yielded an R^2 of .45 with a RMSE of approximately $8,089,273 (Figure 2.21) Feature importance rankings indicated that PDSI was the most influential predictor, with wheat prices and potential evapotranspiration as second and third most important (Figure 2.22). Precipitation was the lowest predictor in terms of feature importance. When evaluating model error across training size (n = 348) using a learning curve analysis, we see considerable variation between training and validation error, which indicates that the model may be overfitting, and would potentially benefit from a reduction of complexity, or an increase in observations to improve performance. A comparison of historical vs. predicted insurance loss is visualized in Figure 2.23, for all 24 counties

combined, with overall model predictions underestimating insurance loss, particularly so in the high loss years of 2009 and 2015.

2.5. Conclusions

The results of our climate lagged correlation analysis provide an interesting view of the spatial and temporal relationships of climate with localized insurance loss, with a particular focus on the region's dryland wheat production. In particular, our results indicate the importance in understanding the varying temporal effects of drought-related climatic variables, as they vary spatially, which is also supported by the previous work of Semenov (2009) and Lobell et al. (2015). As noted previously, we see expected spatial patterns of correlation that align with climate variables within our iPNW study area: for example, we see that counties in the high desert regions of northeastern Oregon, which typically experience very low precipitation rates, were more highly correlated with wheat/drought insurance loss, which may indicate the relative regional sensitivity of these particular counties to precipitation (or lack thereof). Similarly, the higher correlation of PDSI with wheat/drought loss in the most western counties of the region, may indicate that insurance loss in these locations are more sensitive to a composite water-balance context that incorporates temperature, PET, and the available water capacity of soils.

The results of our random forest model efforts indicate the potential importance of economic factors not accounted for in the current model, particularly given the underprediction of years with extreme drought claim totals. Given the considerable declines in wheat prices from 2008 to 2009, during the U.S. great recession (Fan et al., 2016), the underprediction of wheat insurance loss totals in 2009 may indicate the difficulty in the model to factor in economic

considerations, particularly given the possible inflation of drought loss due to economics rather than climatic outcomes. By contrast, in 2015, which was a year of noted severe drought in the iPNW (Mote et al., 2016), the model prediction was considerably closer to the observed loss. The comparisons of these two years may suggest a broader question around insurance loss and the interaction of climate and economics: are there particular extreme economic thresholds which may induce more fraudulent efforts to make climatic-associated loss claims, when, in fact, no such scenario exists?

Given the aforementioned associations of insurance loss with climate change, there are important considerations to contemplate. As future climatic condition in the iPNW will likely lead to increased evapotranspiration rates, increased temperatures, and thus more extreme soil moisture deficits (Suyker & Verma, 2009; United States National Assessment Synthesis Team, 2001), crop insurance programs will likely be considerably negatively impacted. With limited long-term resilience, crop insurance efforts face added financial pressure under prolonged extreme weather conditions: subsidization mechanisms by the federal government to indemnify authorized crop insurance providers have typically only one year of coverage in cases of severe claim payouts (USDA RMA, 2011). As such, future extreme conditions for crop commodity systems, particularly around water availability, will likely increase the likelihood of financial stress with consecutive year drought events. Aspects of insurance loss prediction, as well as understanding potential conditional relationship thresholds (climatic and economic) will become more important, particularly given the need to extend risk management financial tools to areas of the world currently not appropriately indemnified (e.g. Africa). As such, effective modeling efforts to predict the risk based on extreme climatic

conditions, as well as related socio-economic indicators, are becoming paramount (Mann et al., 2019).

An interesting component that was not explored as part of this research is the overlapping and interactive effects of drought combined with heat, and the manifestation of such impacts on overall wheat insurance loss. While agricultural drought implies a reduction in water supply to agronomic systems that may impede or hinder growth rates (and thus reduce yields), heat impacts have the potential to accelerate overall phenological development, which may result in early maturation, lower yields, and potential increases in insurance loss – particularly in cereal systems (Lesk et al., 2016). Understanding these competing and/or countermanding relationships could assist in developing a more accurate representation of climatic impacts on insurance loss, particularly in differing regions of the world. Another extension of this research could employ the use of other drought-associated climatic variables that are supported in other recent research, including: the normalized difference vegetation index (NDVI), precipitation-associated indices, such as the standardized precipitation evaporation index (SPEI), the evaporative demand drought index (EDDI), or simulated soil moisture (Gao et al, 2007; Ahmad et al., 2010). Additional climatic inputs may provide a more enhanced view of insurance loss, particularly related to drought, as well as assisting in separating the differing influences of climate and economics.

2.6. References

Abatzoglou, J. T. (2013). Development of gridded surface meteorological data for ecological applications and modelling. International Journal of Climatology, 33(1), 121–131.

Abatzoglou, J. T., Rupp, D. E., Mote, P. W. (2014). Seasonal climate variability and change in the Pacific Northwest of the United States. Journal of Climate, 27(5), 2125–2142. https://doi.org/10.1175/JCLI-D-13-00218.1

Ahmad, S., Kalra, A., & Stephen, H. (2010). Estimating soil moisture using remote sensing data: A machine learning approach. Advances in Water Resources, 33(1), 69–80. https://doi.org/10.1016/j.advwatres.2009.10.008

Asseng, S., Ewert, F., Rosenzweig, C., Jones, J. W., Hatfield, J. L., Ruane, A. C., … Wolf, J. (2013). Uncertainty in simulating wheat yields under climate change. Nature Climate Change, 3(9), 827–832. https://doi.org/10.1038/nclimate1916

Backlund, P., Janetos, A., Schimel, D. (2008). The effects of climate change on agriculture, land resources, water resources, and biodiversity in the United States. Synthesis and Assessment Product 4.3. Washington, DC: U.S. Environmental Protection Agency, Climate Change Science Program. 240 p.

Barlow, K. M., Christy, B. P., O'Leary, G. J., Riffkin, P. A., & Nuttall, J. G. (2015). Simulating the impact of extreme heat and frost events on wheat crop production: A review. Field Crops Research, 171, 109–119. https://doi.org/10.1016/j.fcr.2014.11.010

Behrens, J. T. (2000). Exploratory data analysis. In Encyclopedia of psychology. Edited by A. E. Kazdin, 303–305. New York: Oxford Univ. Press.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. Taylor & Francis.

Breiman, L. (2001) Random forests. Machine Learning, 45, 5-32. doi:10.1023/A:1010933404324

Chi, J., Waldo, S., Pressley, S. N., Russell, E. S., O'Keeffe, P. T., Pan, W. L., Lamb, B. K. (2017). Effects of Climatic Conditions and Management Practices on Agricultural Carbon and Water Budgets in the Inland Pacific Northwest USA. Journal of Geophysical Research: Biogeosciences, 122(12), 3142–3160. https://doi.org/10.1002/2017JG004148

Clark L.A., Pregibon D. (1992). Tree-based models. In: Chambers JM, Hastie TJ, Eds. Statistical models in S. Pacific Grove (CA): Wadsworth, p 377–419.

Deschênes, B. O., Greenstone, M. (2018). American Economic Association The Economic Impacts of Climate Change : Evidence from Agricultural Output and Random Fluctuations in Weather : Reply Author (s): Olivier Deschênes and Michael Greenstone Source : The American Economic Review , Vol . 102 , , 102(7), 3761–3773.

Du, X., Feng, H., & Hennessy, D. A. (2017). Rationality of choices in subsidized crop insurance markets. American Journal of Agricultural Economics, 99(3), 732–756. https://doi.org/10.1093/ajae/aaw035

Epstein, P. (2015). "Climate Change and Human Health." The New England Journal of Medicine. 353.14 (2005): 1433-6. Web.

Fan, M., Pena, A. A. & Perloff, J. M. (2016). Effects of the Great Recession on the U.S. Agricultural Labor Market. American Journal of Agricultural Economics, 98(4), 1146–1157. https://doi.org/10.1093/ajae/aaw023

FAO. (2016). Climate change and food security: Risks and responses. Nations, Food and Agriculture Organization of the United States. https://doi.org/10.1080/14767058.2017.1347921

Gao, H., Wood, E.F., & Drusch, M. (2007). Copula derived observation operators for assimilating TMI and AMSR-E soil moisture into land surface models. J. Hydrometeorol., 8 (2007), pp. 413-428

Grotjahn, R., Holden, P., Izaurralde, C., Mader, T., & Marshall, E. (2014). Climate Change Impacts in the U.S. - Agriculture, 150–174. https://doi.org/10.7930/J02Z13FR.On

Hatfield, J. & Dold, C. (2018). Climate Change Impacts on Corn Phenology and Productivity. Chapter 6, Corn-Production and Human Health in Changing Climate. http://dx.doi.org/10.5772/57353

Hatfield, J., G. Takle, R. Grotjahn, P. Holden, R. C. Izaurralde, T. Mader, E. Marshall, & D. Liverman (2014). Ch. 6: Agriculture. Climate Change Impacts in the United States: The Third National Climate Assessment, J. M. Melillo, Terese (T.C.) Richmond, and G. W. Yohe, Eds., U.S. Global Change Research Program, 150-174. doi:10.7930/J02Z13FR.

Howitt R., D. MacEwan, J. Medellin- Azuara, J.R. Lund, & D.A. Sumner. (2015). "Preliminary Analysis: 2015 Drought Economic Impact Study." Center for Watershed Sciences, UC Davis, California. https:// watershed.ucdavis.edu/files/content/ news/2015Drought_PrelimAnalysis.pdf

Karimi, T., Stöckle, C. O., Higgins, S. S., Nelson, R. L. & Huggins, D. (2017). Projected Dryland Cropping System Shifts in the Pacific Northwest in Response to Climate Change. Frontiers in Ecology and Evolution, 5(April), 1–9. https://doi.org/10.3389/fevo.2017.00020

Lesk, C., Rowhani, P., & Ramankutty, N. (2016). Influence of extreme weather disasters on global crop production. Nature, 529, 84. Retrieved from https://doi.org/10.1038/nature16467

Lobell, D. B., Schlenker, W., & Costa-Robert, J. (2011). Climate trends and global crop production since 1980. Science, 333(2011), 616–620. https://doi.org/10.1126/science.1204531

Lobell, D. B., Hammer, G. L., Chenu, K., Zheng, B., Mclean, G., & Chapman, S. C. (2015). The shifting influence of drought and heat stress for crops in northeast Australia. Global Change Biology, 21(11), 4115–4127. https://doi.org/10.1111/gcb.13022

Manandhar, S., Pandey, V. P., Kazama, F., & Kazama, S. (2014). Economics of climate change. In Climate Change and Water Resources. https://doi.org/10.1201/b16969

Mankin, J. S., N.S. Diffenbaugh, (2015). Influence of temperature and precipitation variability on near-term snow trends, Climate Dynamics, 45 1099-1116, DOI 10.1007/s00382-014-2357-4.

Marlier, M. E., Xiao, M., Engel, R., Livneh, B., Abatzoglou, J. T., & Lettenmaier, D. P. (2017). The 2015 drought in Washington State: a harbinger of things to come? Environmental Research Letters, 12(11), 114008. https://doi.org/10.1088/1748-9326/aa8fde

Mann, M.L. Warner, J.M. & Malik, A.S. (2019). Predicting high-magnitude, low-frequency crop losses using machine learning: An application to cereal crops in Ethiopia. Climatic Change 154(1-2): 211–227. https://doi.org/10.1007/s10584-019-02432-7Manandhar, S., Pandey, V. P., Kazama, F. & Kazama, S. (2014). Economics of climate change. In Climate Change and Water Resources. https://doi.org/10.1201/b16969

Mote, P., Abatzoglou, J. & Kunkel, K. (2013). Climate change in the Northwest, in Climate Change in the Northwest: Implications for Our Landscapes, Waters, and Communities, edited by M. Dalton, P. W. Mote, and A. K. Snover, chap. 2, 224 pp., Island Press, Washington D. C.

Mote, P. W., Rupp, D. E., Li, S., Sharp, D. J., Otto, F., Uhe, P. F., . . . Allen, M. R. (2016). 2015 snowpack in the western United States. Geophysical Research Letters, 1–9. http://doi.org/10.1002/2016GL069965

Nelson, G. C., Valin, H., Sands, R. D., Havlík, P., Ahammad, H., Deryng, D., & Willenbockel, D. (2014). Climate change effects on agriculture: Economic responses to biophysical shocks. Proceedings of the National Academy of Sciences, 111(9), 3274 LP – 3279. https://doi.org/10.1073/pnas.1222465110

Palmer, W. C. (1965). Research Paper no. 45. US Department of Commerce.

Prasad, A. M., Iverson, L. R. & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems, 9(2), 181–199. https://doi.org/10.1007/s10021-005-0054-1

Perlich, C., Provost, F. & Simonoff, J. S. (2004). Tree induction vs. Logistic regression: A learning-curve analysis. Journal of Machine Learning Research, 4(2), 211–255. https://doi.org/10.1162/153244304322972694

Quinlan, J. (1986). Strategic Induction of Decision Trees BT - Research and Development in Expert Systems XV. Research and Development in Expert Systems XV, 1(Chapter 2), 15–26. https://doi.org/10.1023/A:1022643204877

Redmond, K. (2002). the Depiction of Drought. Bams, (December), 1143–1147. https://doi.org/10.1175/1520-0477-83.8.1143

Rezaei, E. E., Siebert, S., Hüging, H. & Ewert, F. (2018). Climate change effect on wheat phenology depends on cultivar change. Scientific Reports, 8(1), 4891. https://doi.org/10.1038/s41598-018-23101-2

Roesch-McNally, G. (2018). U.S. Inland Pacific Northwest Wheat Farmers' Perceived Risks: Motivating Intentions to Adapt to Climate Change? Environments, 5(4), 49. https://doi.org/10.3390/environments5040049

Rosenzweig, C., Iglesias, A., Yang, X. B., Epstein, Paul, R. & Chivian, E. (2001). Climate change and extreme weather events: Implication for food production, plant diseases, and pests. Global Change & Human Health. https://doi.org/10.1007/s10584-010-9834-5

Rosenzweig, C., Martin, B., & Parry, L. (1994). Potential impact of climate change on world food supply.

Sacks W. J., & Kucharik C J. (2011). Crop management and phenology trends in the U.S. Corn Belt: Impacts on yields, evapotranspiration and energy balance. Agricultural and Forest Meteorology, 151(7): 882-894.

Sandison, D. I. (2017). 2015 Drought and Agriculture, 495 (February). Retrieved from https://agr.wa.gov/FP/Pubs/docs/495-2015DroughtReport.pdf

Schillinger, W. F., Papendick, R. I., & Mccool, D. K. (2010). Soil and Water Challenges for Pacific Northwest Agriculture. Soil and Water Conservation Advances in the United States, 47–80.

Schlenker, W. & Roberts, M. J. (2009). Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change. Proceedings of the National Academy of Sciences, 106(37), 15594–15598. https://doi.org/10.1073/pnas.0906865106

Seamon, E., Gessler, P.E., Abatzogou, J.T., Mote, P.W., & Lee, S.S. (2019a). Agricultural insurance loss analysis of the Pacific Northwest, USA: 2001 – 2015. Doctoral dissertation chapter 1, University of Idaho.

Semenov, M. A. (2009). Impacts of climate change on wheat in England and Wales. Journal of the Royal Society Interface, 6(33), 343–350. https://doi.org/10.1098/rsif.2008.0285

Singh, R. B. (2012). Climate Change and Food Security. In Improving Crop Productivity in Sustainable Agriculture. https://doi.org/10.1002/9783527665334.ch1

Stöckle, C. O., Nelson, R. L., Higgins, S., Brunner, J., Grove, G., Boydston, R., … Kruger, C. (2010). Assessment of climate change impact on Eastern Washington agriculture. Climatic Change, 102(1–2), 77–102. https://doi.org/10.1007/s10584-010-9851-4

Stöckle, C. O., Higgins, S., Nelson, R., Abatzoglou, J., Huggins, D., Pan, W., … Brooks, E. (2018). Evaluating opportunities for an increased role of winter crops as adaptation to climate change in dryland cropping systems of the U.S. Inland Pacific Northwest. Climatic Change, 146(1–2), 247–261. https://doi.org/10.1007/s10584-017-1950-z

Suyker, A. E., & Verma, S. B. (2009). Evapotranspiration of irrigated and rainfed maize-soybean cropping systems. Agricultural and Forest Meteorology, 149(3–4), 443–452. https://doi.org/10.1016/j.agrformet.2008.09.010

Therneau T. & Atkinson E. (1997). An introduction to recursive partitioning using rpart routines. Mayo Foundation. Available: http://www.mayo.edu/hsr/techrpt/61.pdf.

Trenberth, K. E. (2000). Effects of changing climate on weather and human activities. Sausalito, Calif.: University Science Books.

United States National Assessment Synthesis Team. (2001). Climate change impacts on the United States : The potential consequences of climate variability and change : Foundation ; a report of the national assessment synthesis team ; U.S. global change research program. Cambridge: Cambridge Univ. Press.

USDA Risk Management Agency, (2011). History of the Crop Insurance Program. https://legacy.rma.usda.gov/aboutrma/what/history.html

United States Global Change Research Program. (2017). Climate Science Special Report: Fourth National Climate Assessment, Volume I. (D. J. Wuebbles, D. W. Fahey, K. A. Hibbard, D. J. Dokken, B. C. Stewart, & T. K. Maycock, Eds.). Washington, DC: U.S. Global Change Research Program. https://doi.org/10.7930/J0J964J6

Verbyla D.L. (1987). Classification trees a new discrimination tool. Can J. For Res 17:1150-2

Walker, K., & Rahe, M. (2015). Oregon Agriculture, Food and Fiber: An Economic Analysis. Oregon Department of Agriculture, Oregon State Extension Service, Rural Studies Program (December).

Yorgey, G. & Kruger, C. E. (2017). Advances in Dryland Farming in the Inland Pacific Northwest. Washington State University Extension Publications. Retrieved from https://books.google.com/books?id=vZnEswEACAAJ

Figure 2.1. 24-county inland Pacific Northwest (iNPW) study area, which includes counties from Washington, Idaho, and Oregon. Additionally noted: on the inset map in upper left depicts the three main agricultural regions in the Pacific Northwest which include, in addition to the iPNW: Oregon's Willamette Valley (green), and southern Idaho's Snake River Valley (red).

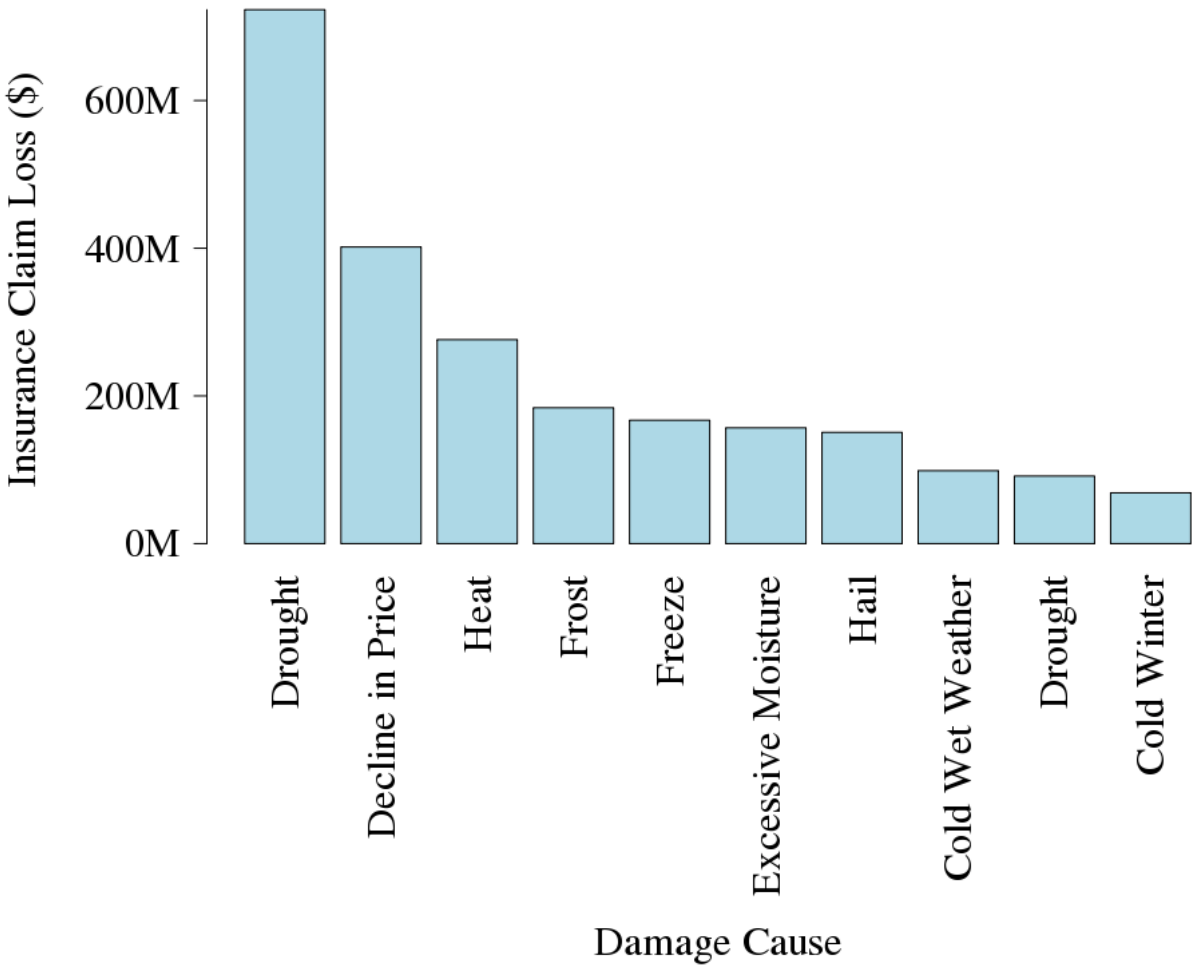


Figure 2.2. Breakdown of damage cause insurance claims across the 24-county iPNW region, from 2001 to 2015, by dollar amount. Drought and heat combined to result in $950 million in total losses.

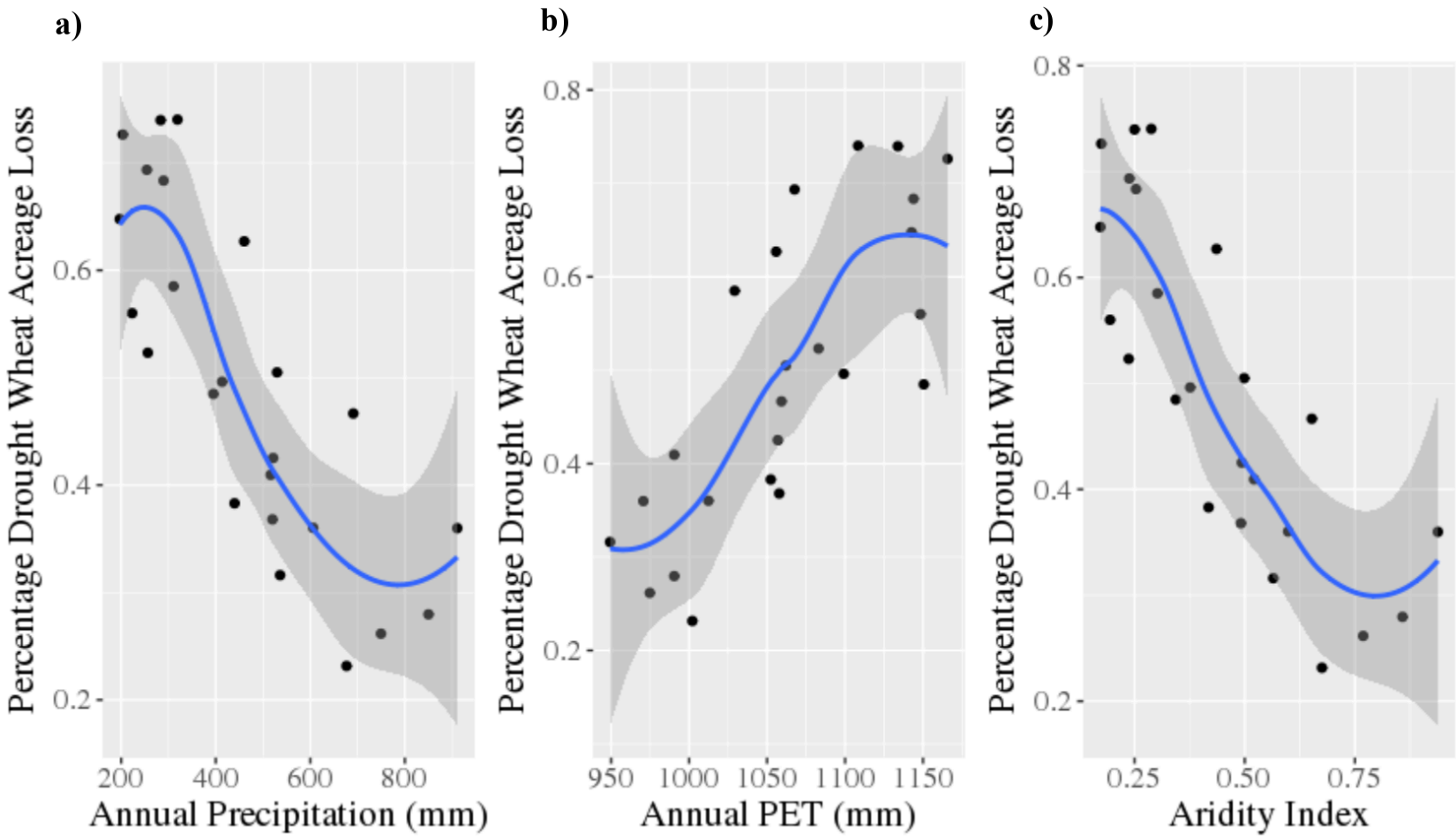
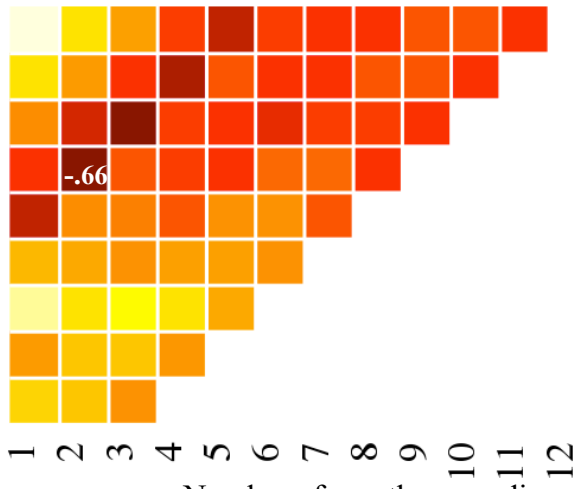


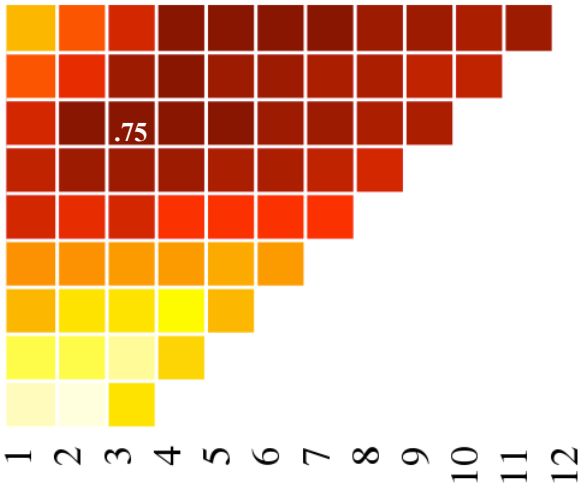
Figure 2.3. County level comparisons of the total % of wheat insurance loss acreage due to drought, from 2001 to 2015, vs (a) annual precipitation totals, (b) annual potential evapotranspiration (PET) totals, and (c) aridity (precipitation divided by potential evapotranspiration). Acreage values by insurance claim are only available after 2000. Each observation represents an individual counties' average annual value, for all years, from 2001 to 2015.



Precipitation vs. Wheat/Drought Loss: Whitman County, WA

April/May/June precipitation has highest correlation with annual insurance loss ($) for Whitman County, WA

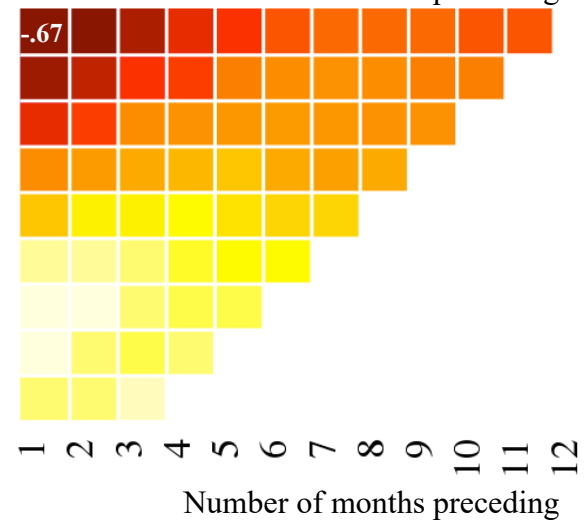
Range: -.16 to -.66



Potential Evapotranspiration vs. Wheat/Drought Loss: Whitman County, WA

Apr/May/June/July PET has highest correlation with annual insurance loss ($) for Whitman County, WA

Range: .24 to .75



PDSI vs. Wheat/Drought Loss: Whitman County, WA

Aug/Sept PDSI has highest correlation with annual insurance loss ($) for Whitman County, WA

Range: -.22 to -.67

Figure 2.4. Example climate and wheat insurance loss correlation matrices for Whitman county, WA, 2001 to 2015, due to drought. Correlation values are absolute (R).

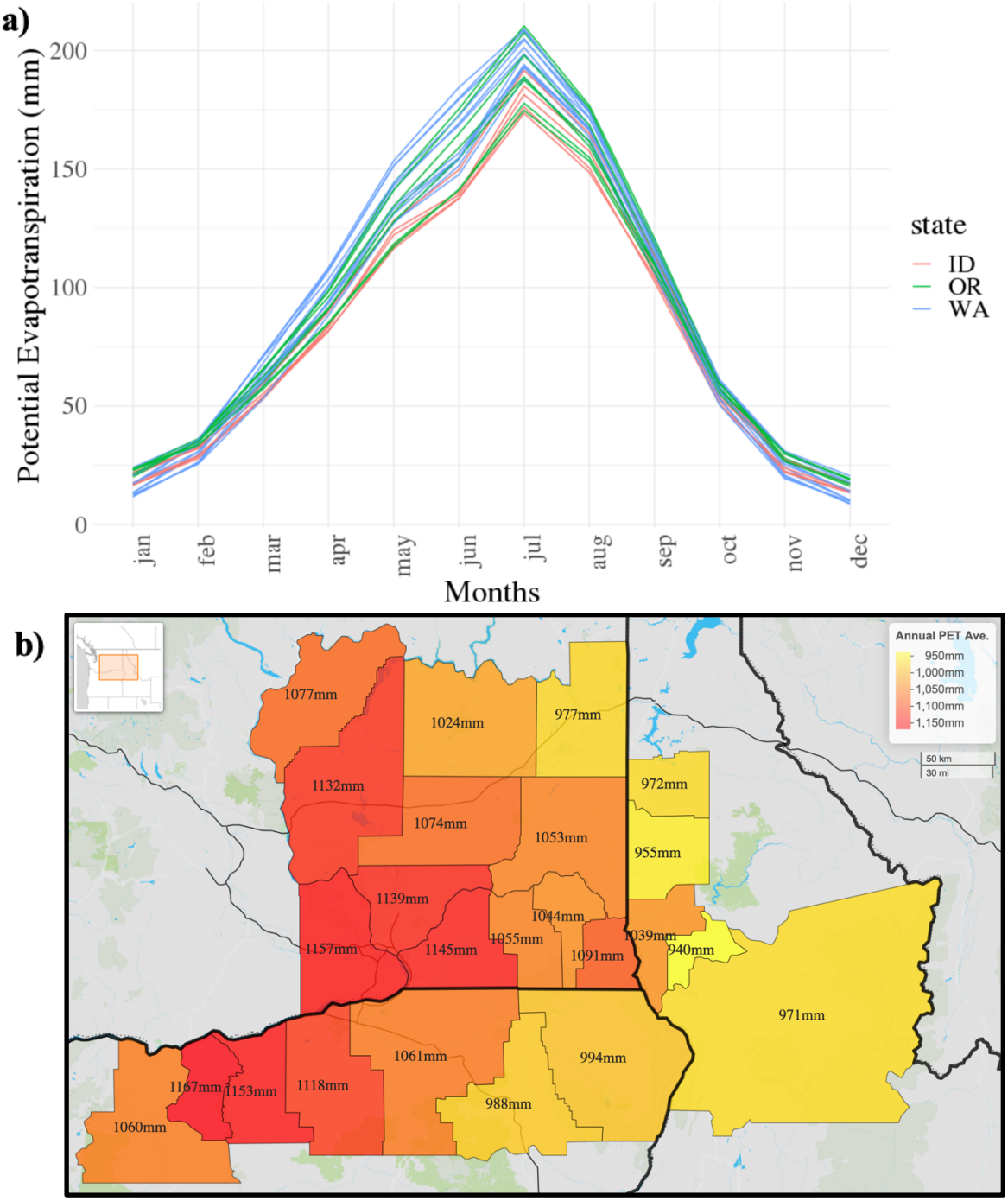


Figure 2.5. Top panel shows (a) potential evapotranspiration (PET) monthly averages per county, from 2001 to 2015, grouped by state. Bottom panel (b) shows annual PET averages (2001 to 2015). Note the spatial trend of PET increasing from east to west.

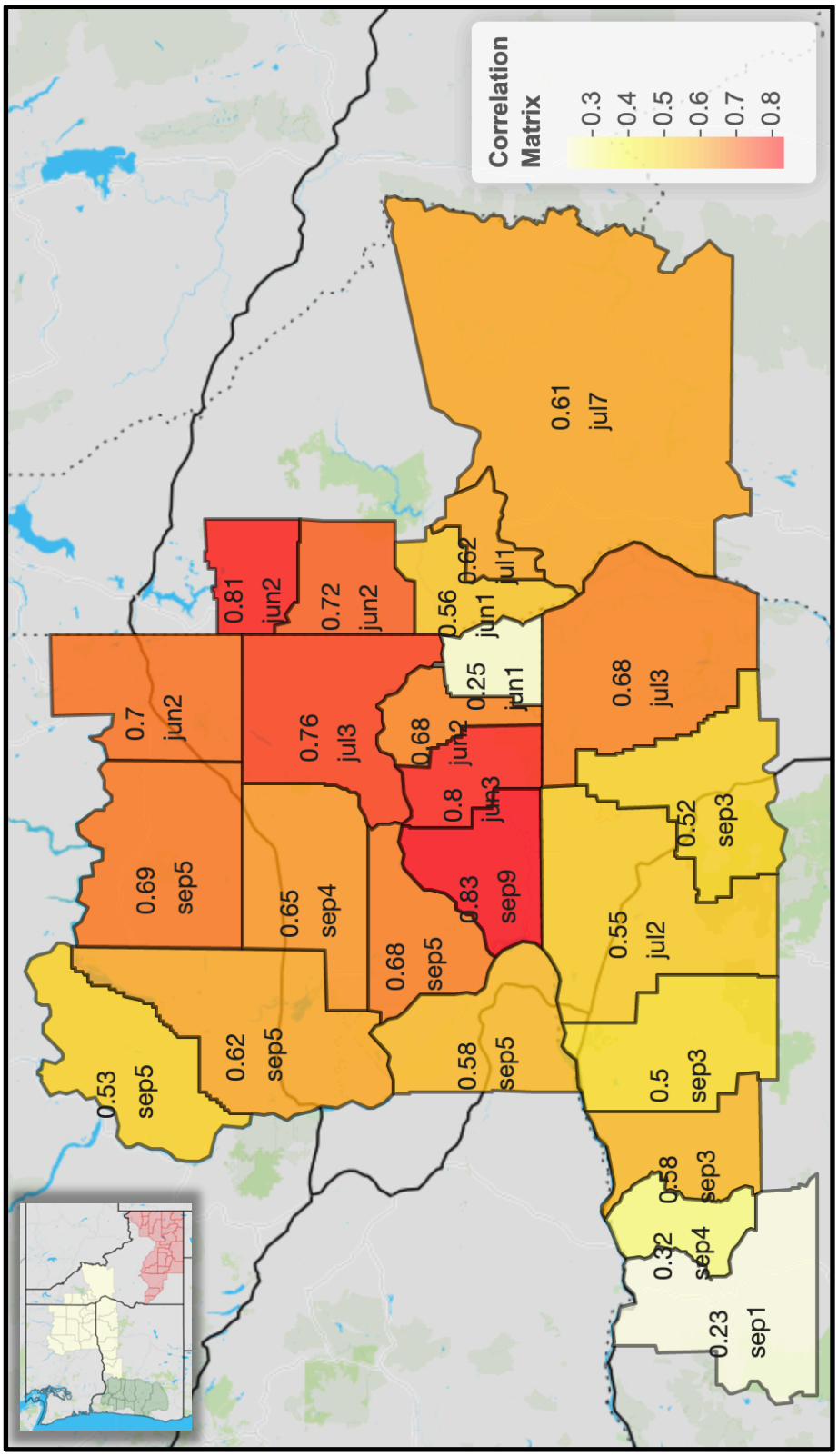


Figure 2.6. Potential evapotranspiration and wheat/drought insurance loss correlations by county, using the optimum monthly combination using data from 2001 to 2015. The highest correlations occur in the eastern/central portions of the study area.

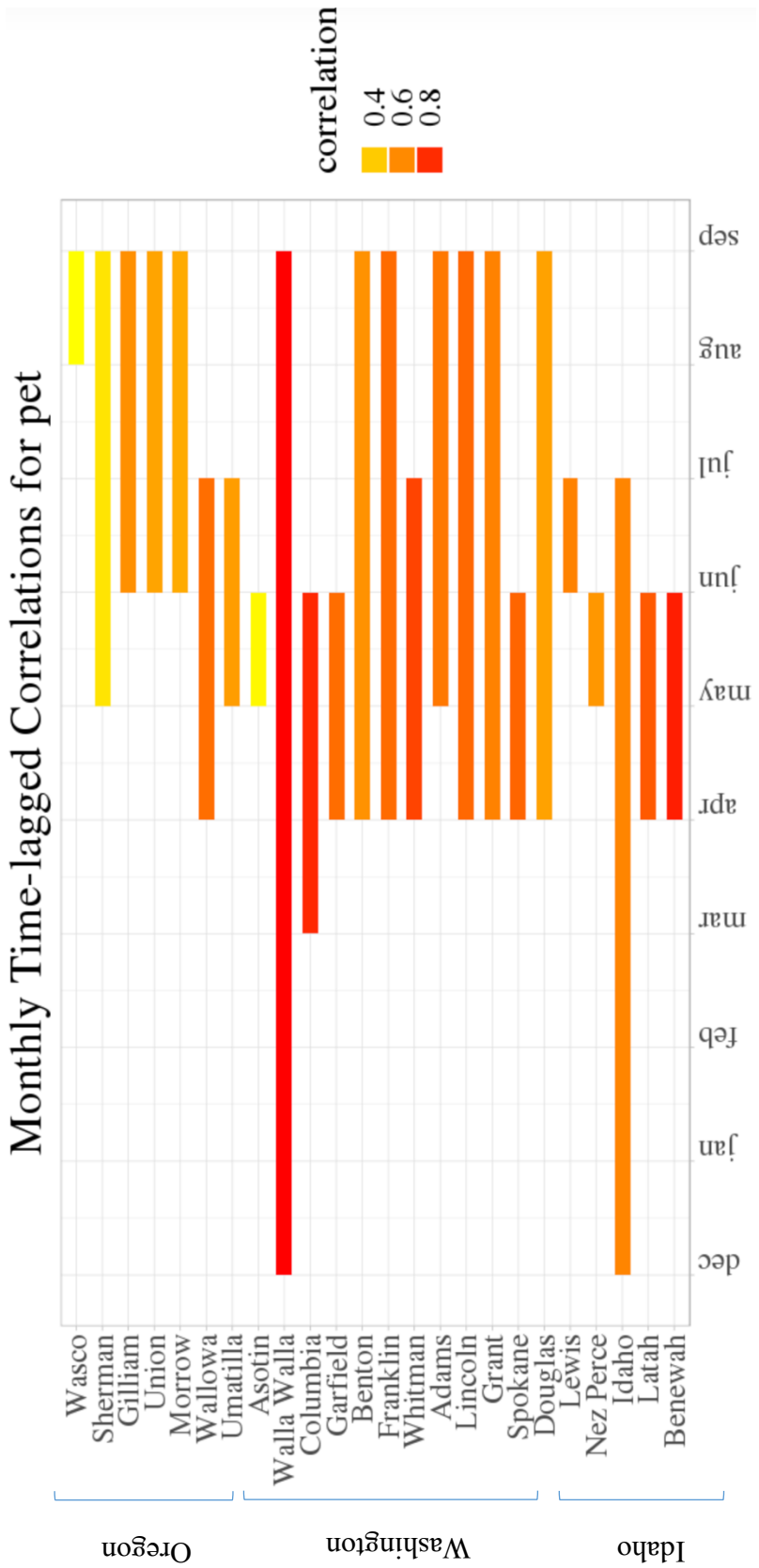


Figure 2.7. Potential evapotranspiration and wheat/drought insurance loss correlations by county, indicating the optimum time windows for each county and the associated correlation value.

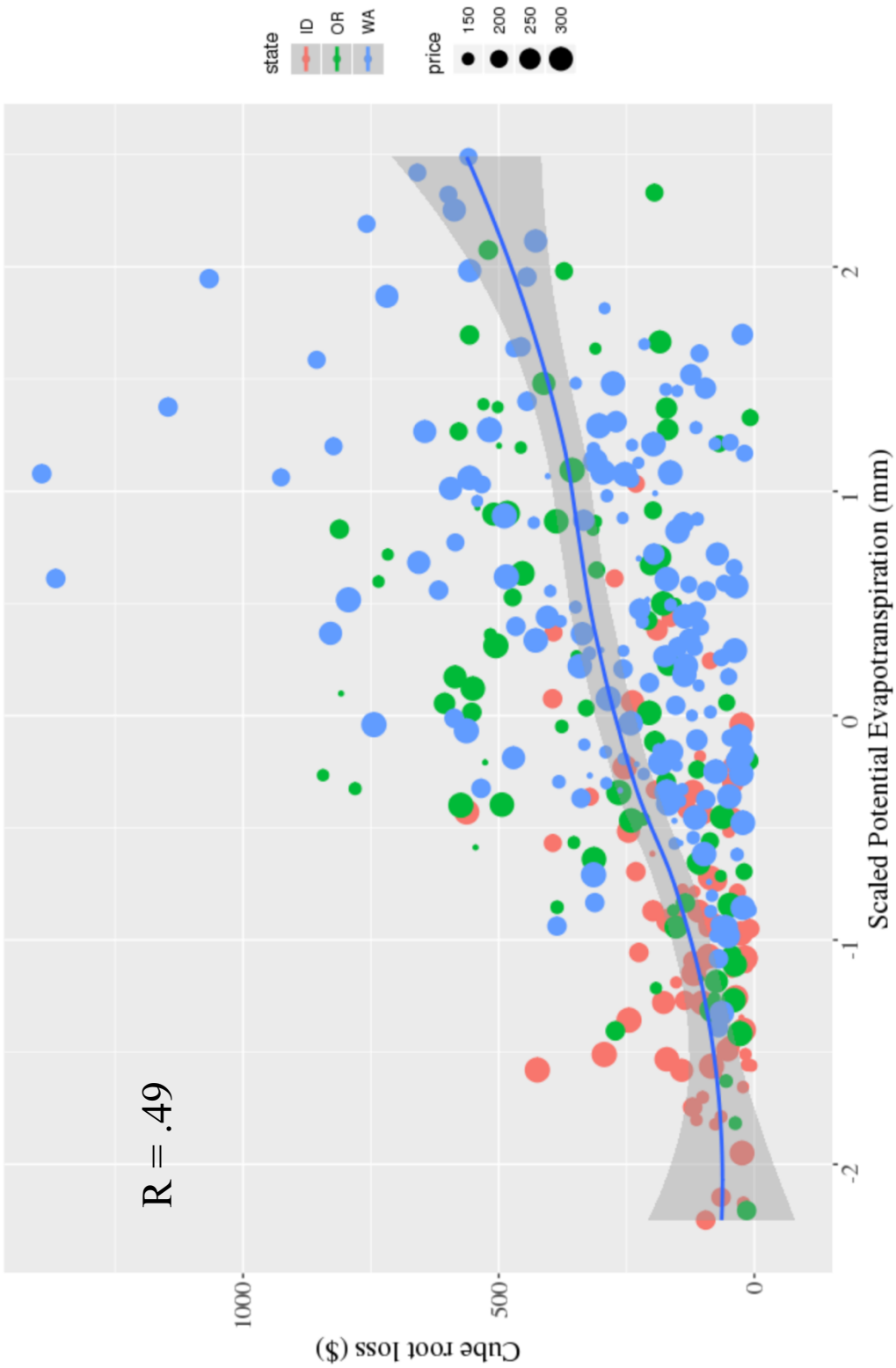


Figure 2.8. Absolute correlation (R) between annual wheat/drought insurance dollar loss (by county), and scaled potential evapotranspiration, which was refined as a result of the time lagged correlation approach (2001 – 2015). The size of observations represents the average price of wheat for that year ($ per metric ton).

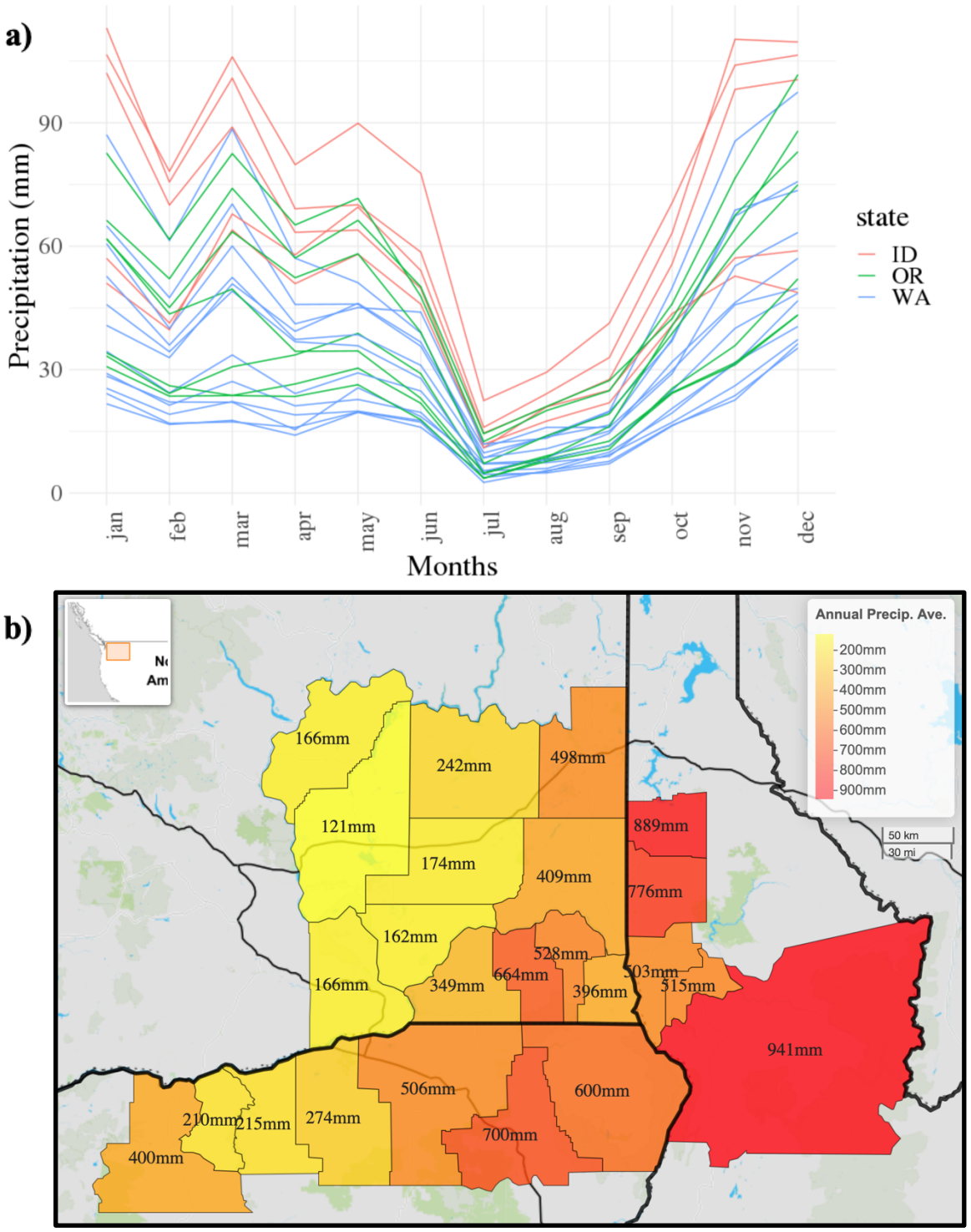


Figure 2.9. Top panel shows (a) precipitation monthly averages per county, from 2001 to 2015, grouped by state. Bottom panel (b) shows annual precipitation averages (2001 to 2015). Note the spatial trend increasing from northwest to southeast.

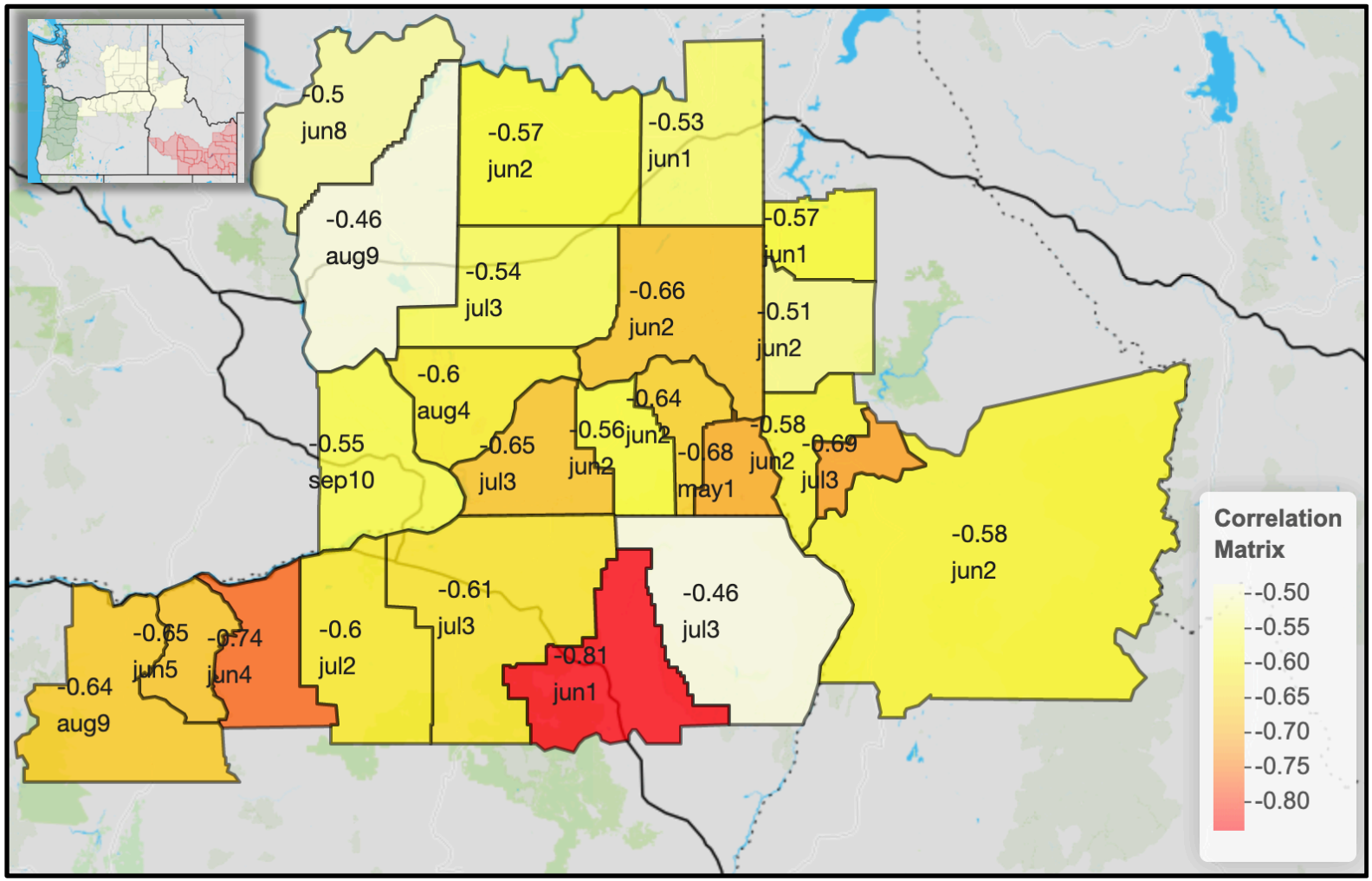


Figure 2.10. Precipitation and wheat/drought insurance loss correlations by county, using the optimum monthly combination using data from 2001 to 2015.

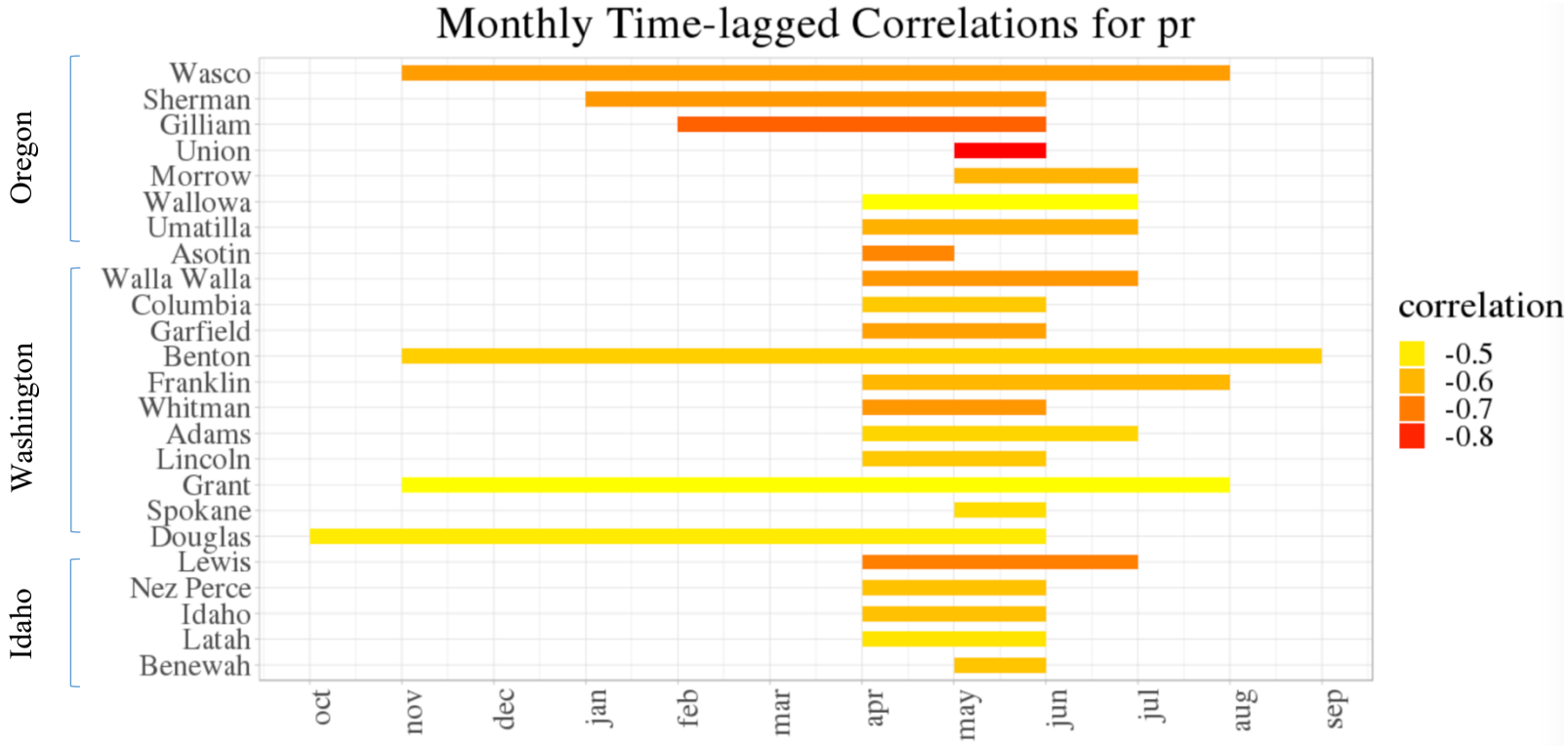


Figure 2.11. Precipitation and wheat/drought insurance loss correlations by county, indicating the optimum time windows for each county and the associated correlation value.

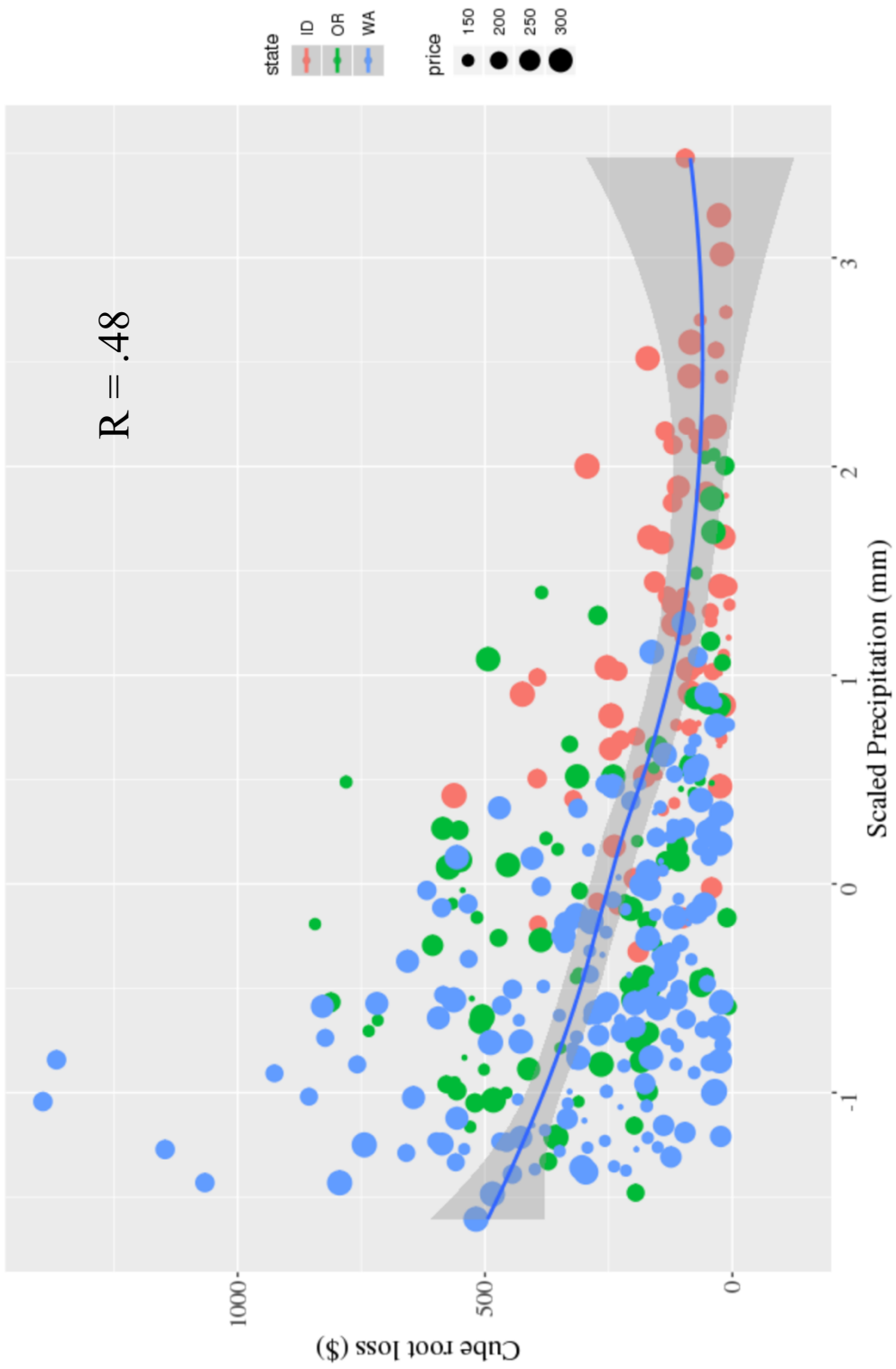


Figure 2.12. Absolute correlation (R) between annual wheat/drought insurance dollar loss (by county), and the zscore of precipitation, which was refined as a result of the time lagged correlation approach (2001 – 2015). The size of observations represents the average price of wheat for that year ($ per metric ton).

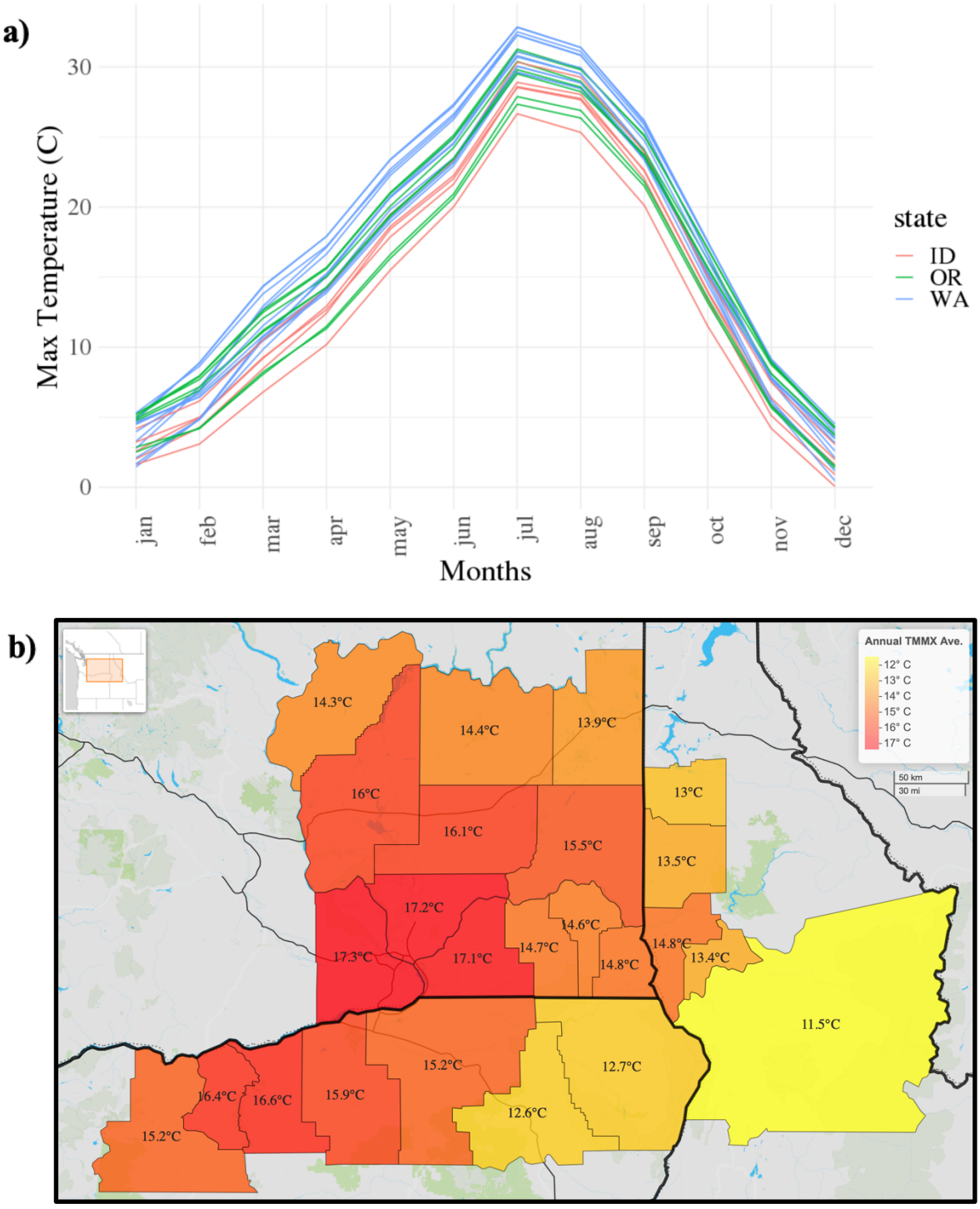


Figure 2.13. Top panel shows (a) max temperature monthly averages per county, from 2001 to 2015, grouped by state. Bottom panel (b) shows annual max temperature averages (2001 to 2015). Note the spatial trend increasing from southeast to northwest.

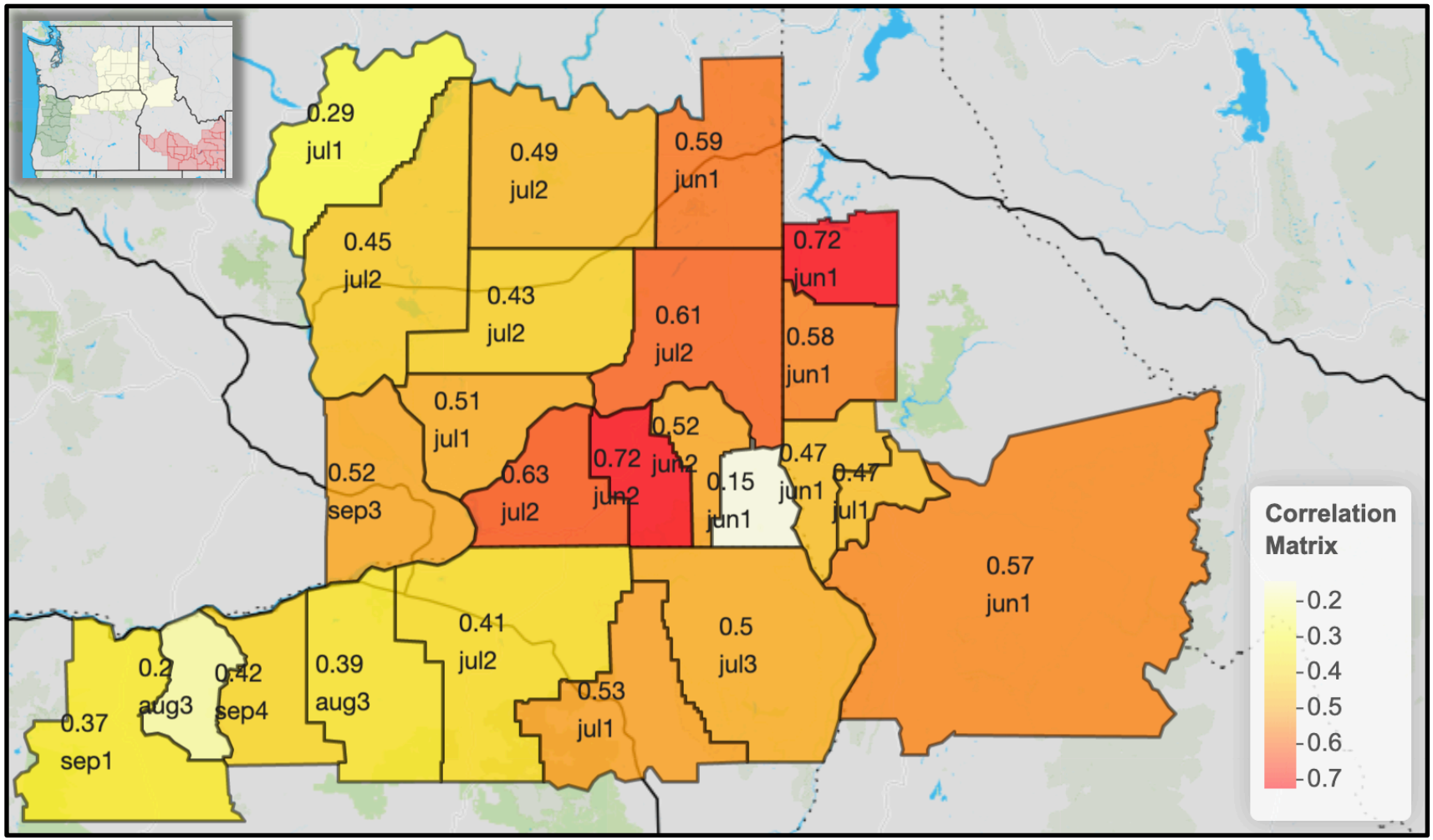


Figure 2.14. Maximum temperature and wheat/drought insurance loss correlations by county, using the optimum monthly combination using data from 2001 to 2015.

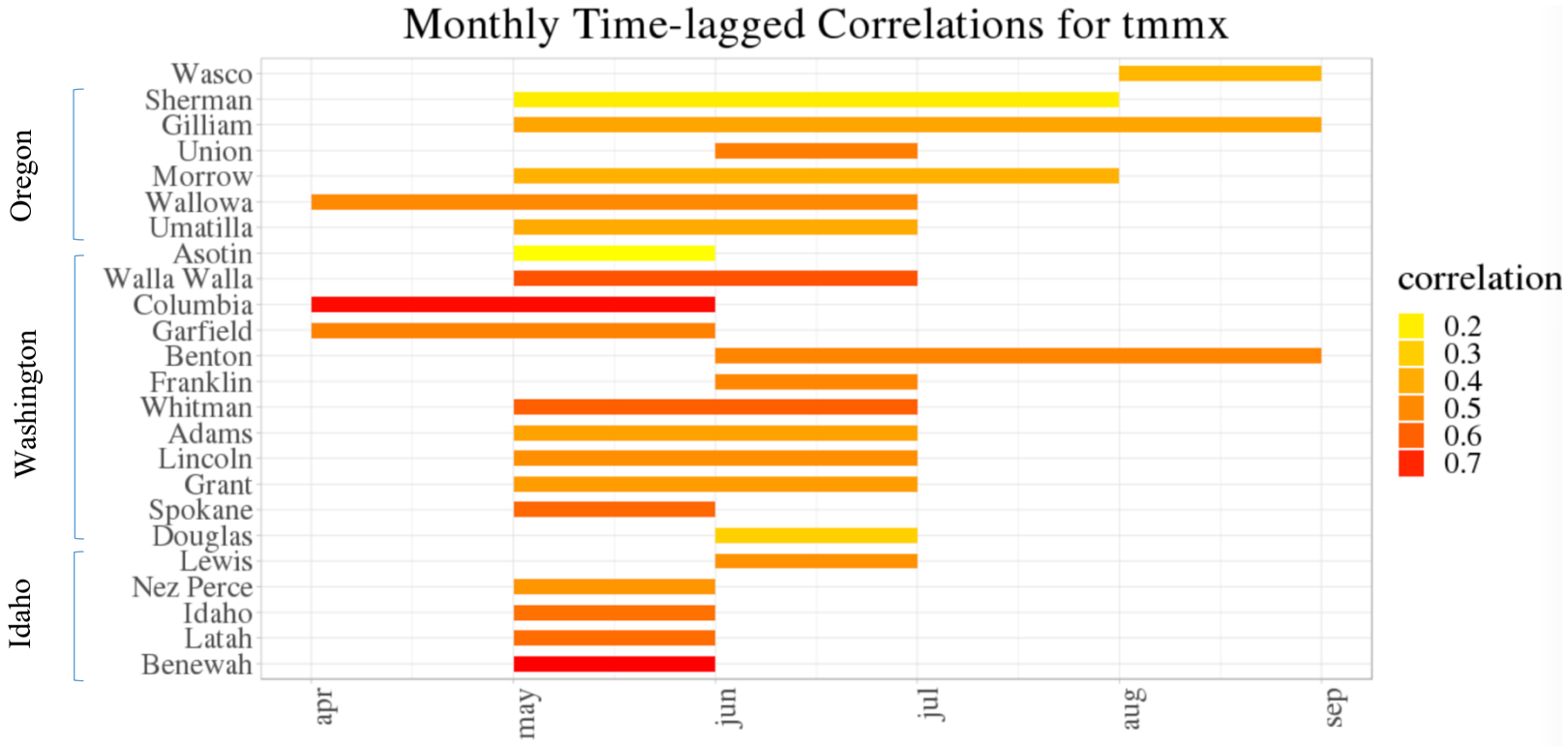


Figure 2.15. Maximum temperature and wheat/drought insurance loss correlations by county, indicating the optimum time windows for each county and the associated correlation value.

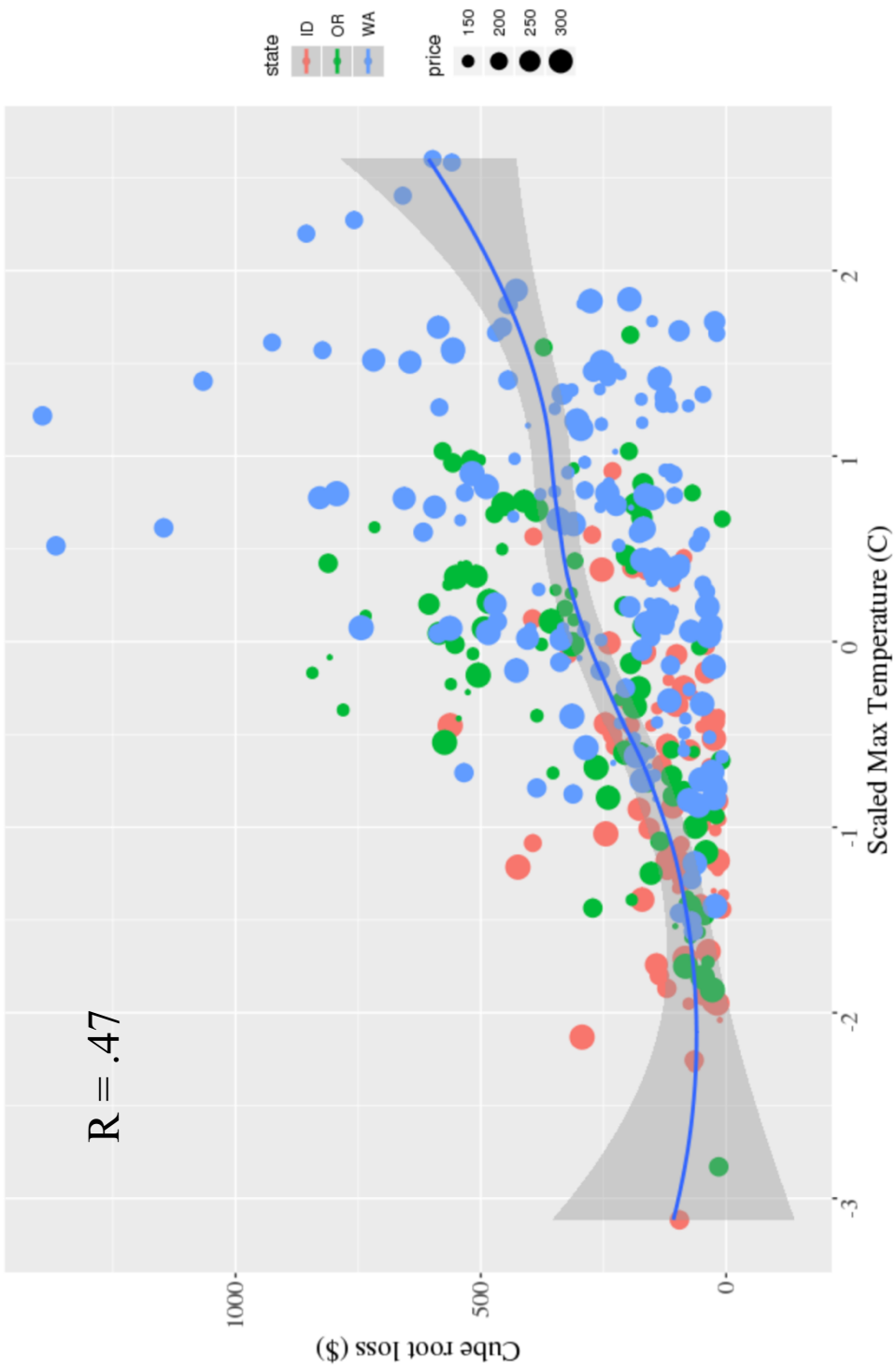


Figure 2.16. Absolute correlation (R) between annual wheat/drought insurance dollar loss (by county), and the zscore of maximum temperature, which was refined as a result of the time lagged correlation approach (2001 – 2015). The size of observations represents the average price of wheat for that year ($ per metric ton)

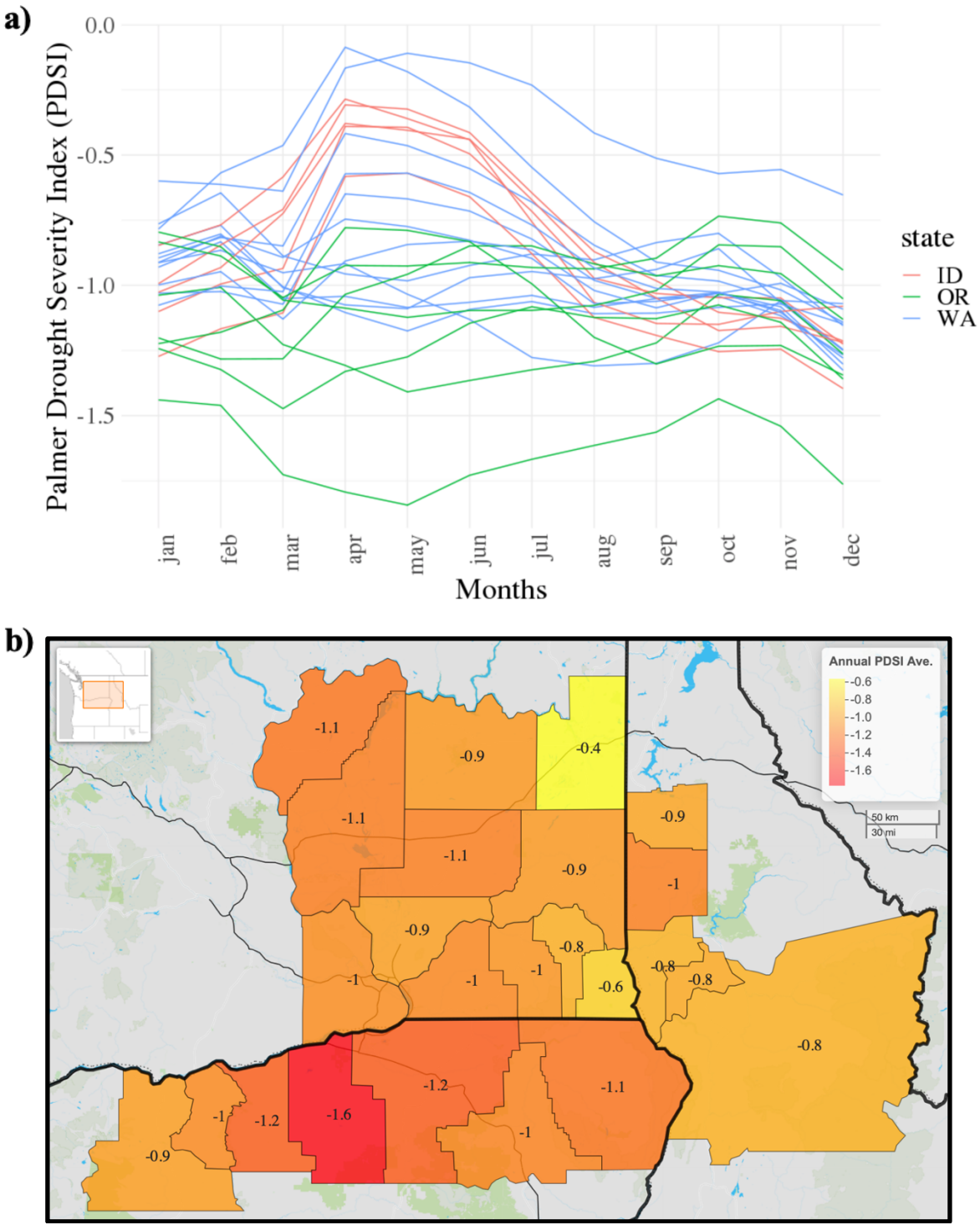


Figure 2.17. Top panel shows (a) PDSI monthly averages per county, from 2001 to 2015, grouped by state. Bottom panel (b) shows annual PDSI averages (2001 to 2015). Note the spatial trend increasing from southeast to northwest.

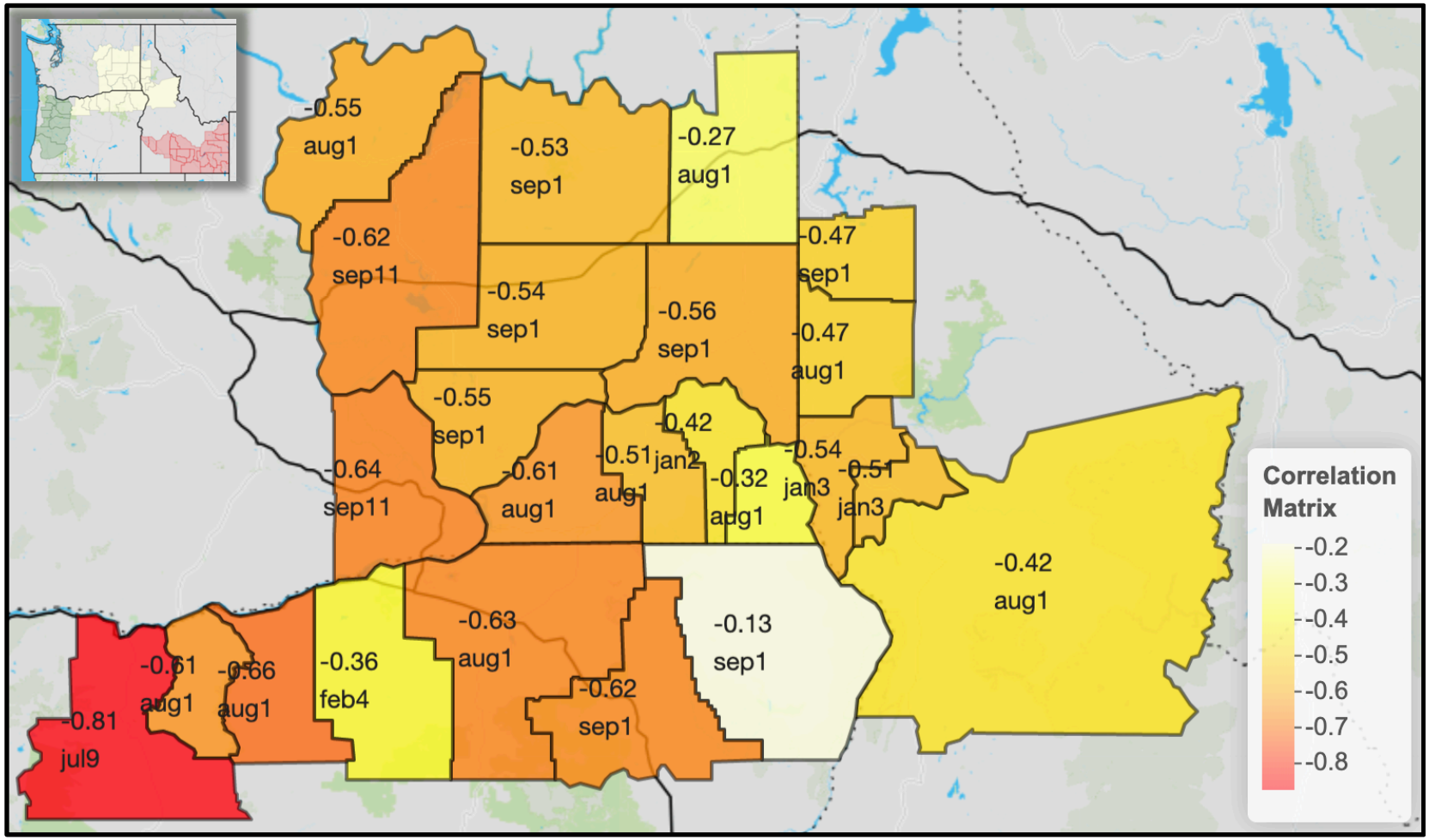


Figure 2.18. Palmer Drought Severity Index (PDSI) and wheat/drought insurance loss correlations by county, using the optimum monthly combination using data from 2001 to 2015. The highest correlations occur in the western portions of the study area.

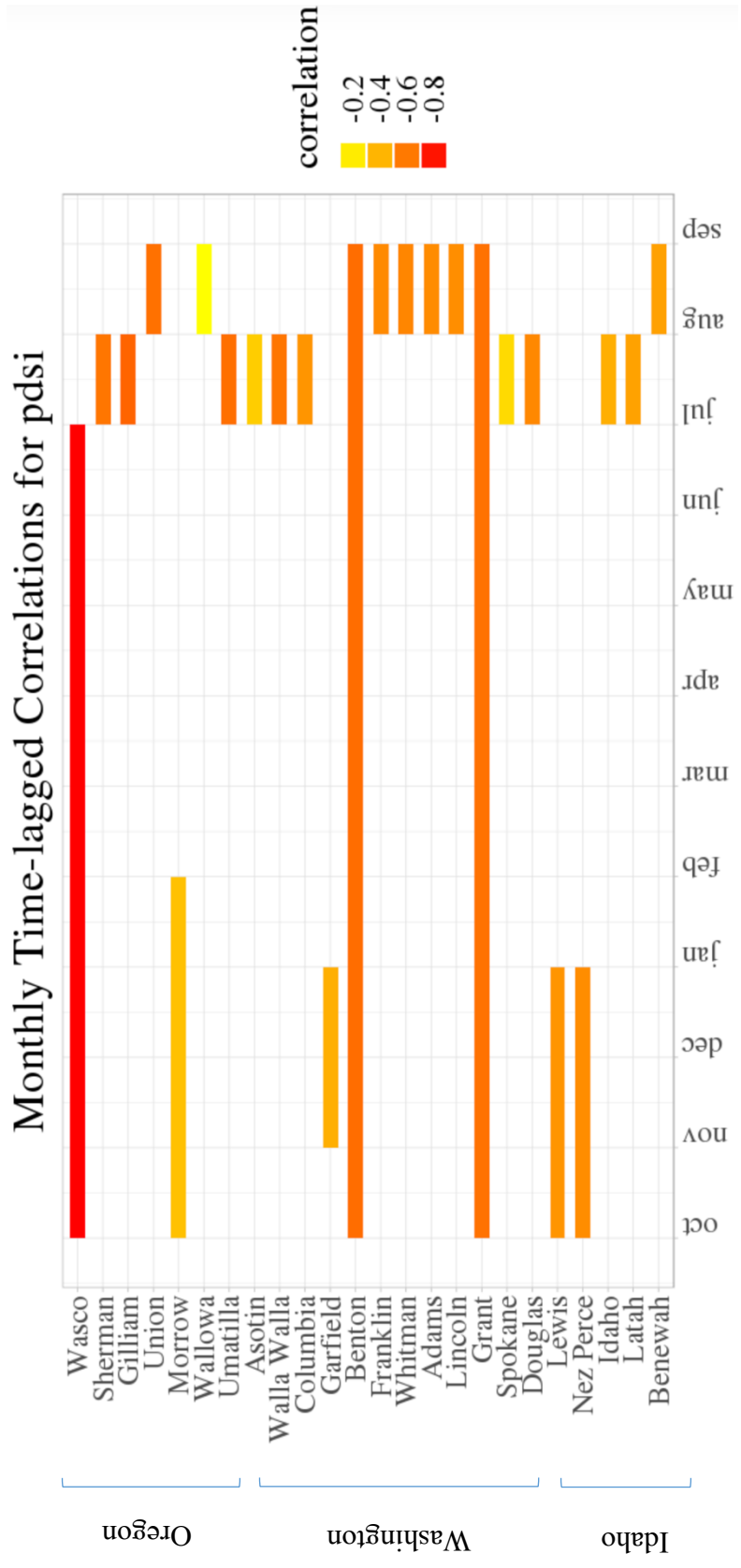


Figure 2.19. Palmer Drought Severity Index (PDSI) and wheat/drought insurance loss correlations by county, indicating the optimum time windows for each county and the associated correlation value.

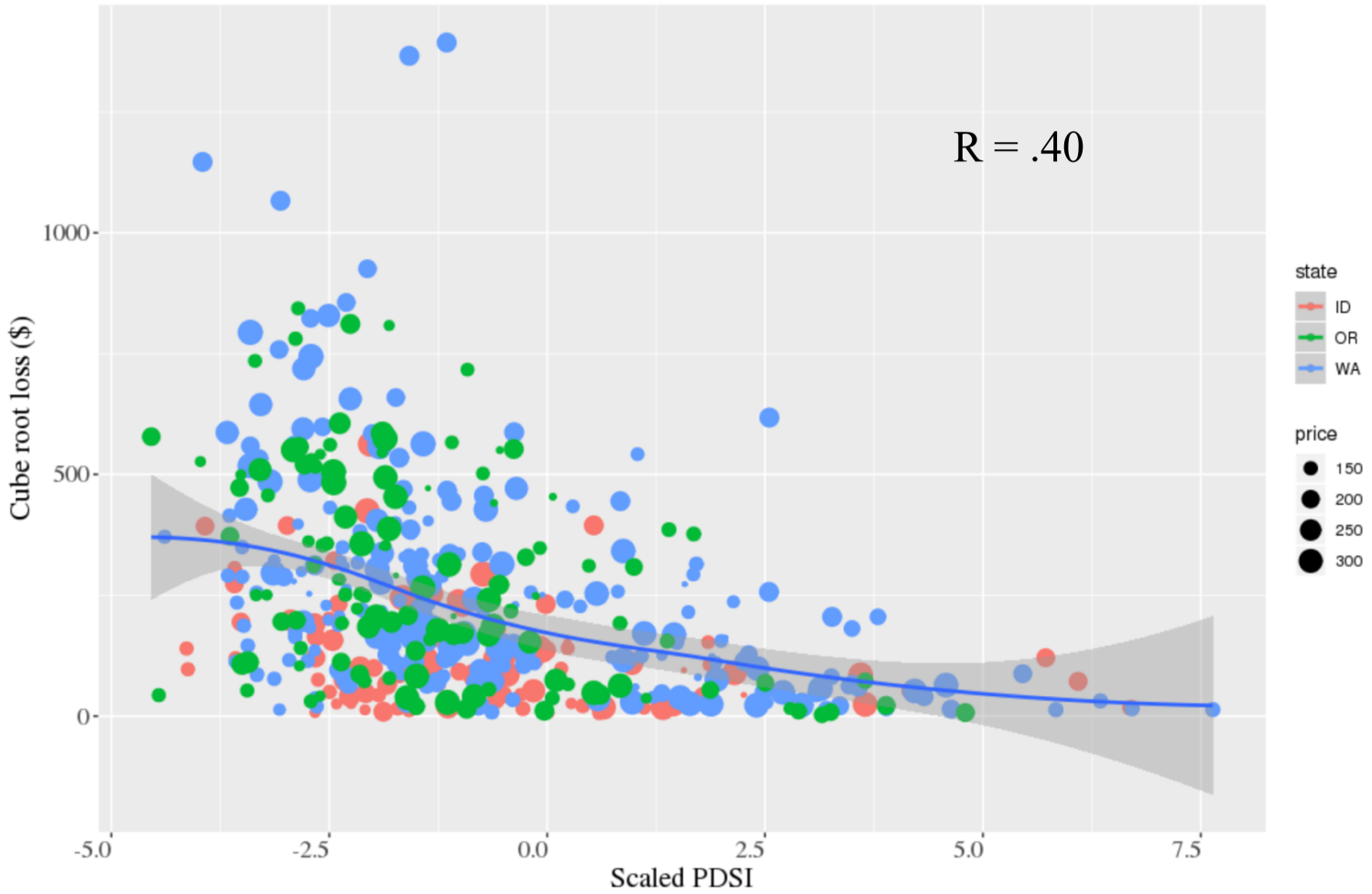


Figure 2.20. Absolute correlation (R) between annual wheat/drought insurance dollar loss (by county), and the zscore of PDSI, which was refined as a result of the time lagged correlation approach (2001 – 2015). The size of observations represents the average price of wheat for that year ($ per metric ton).

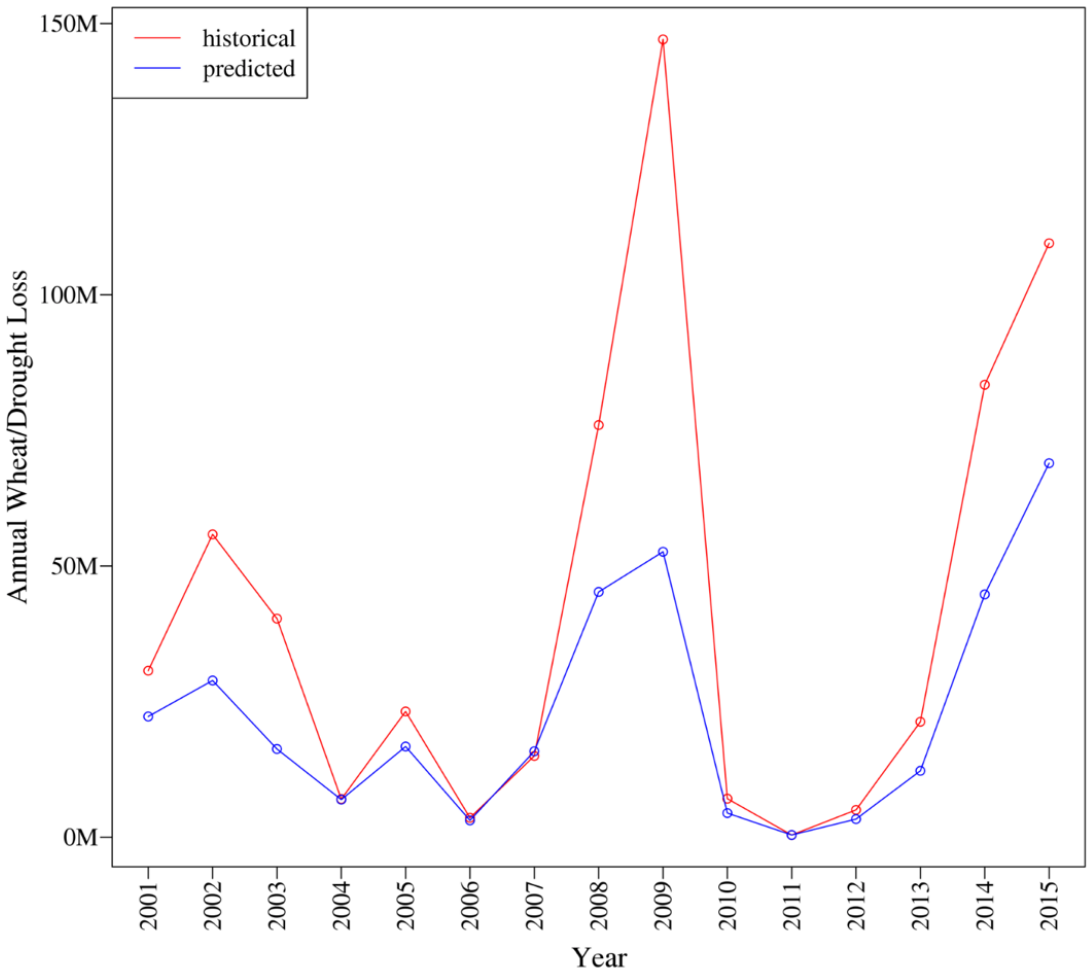


Figure 2.21. Historical vs. predicted annual wheat insurance loss ($) due to drought, constructed using a random forest model (number of trees = 1000), for the 24 county iPNW study area. Input variables were precipitation, maximum temperature, and potential evapotranspiration, as well as annual wheat pricing, from 2001 to 2015. Climate variables were refined using the aforementioned time-lagged correlation methodology (R^2 = .47, RMSE = $8,089,273)

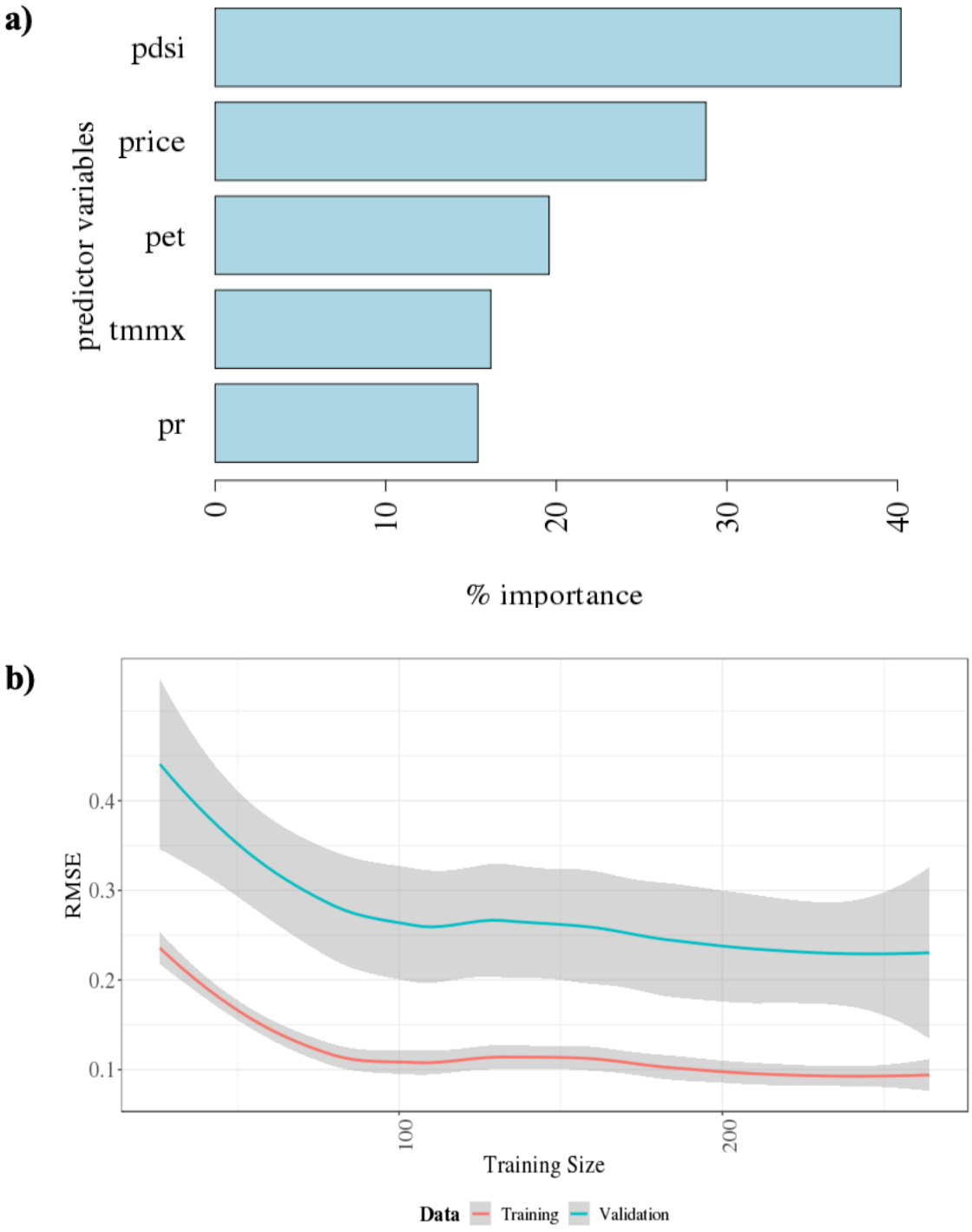


Figure 2.22. (a) Random forest feature importance (total trees = 1000), as well as a (b) learning curve comparison of training dataset error vs. validation dataset error.

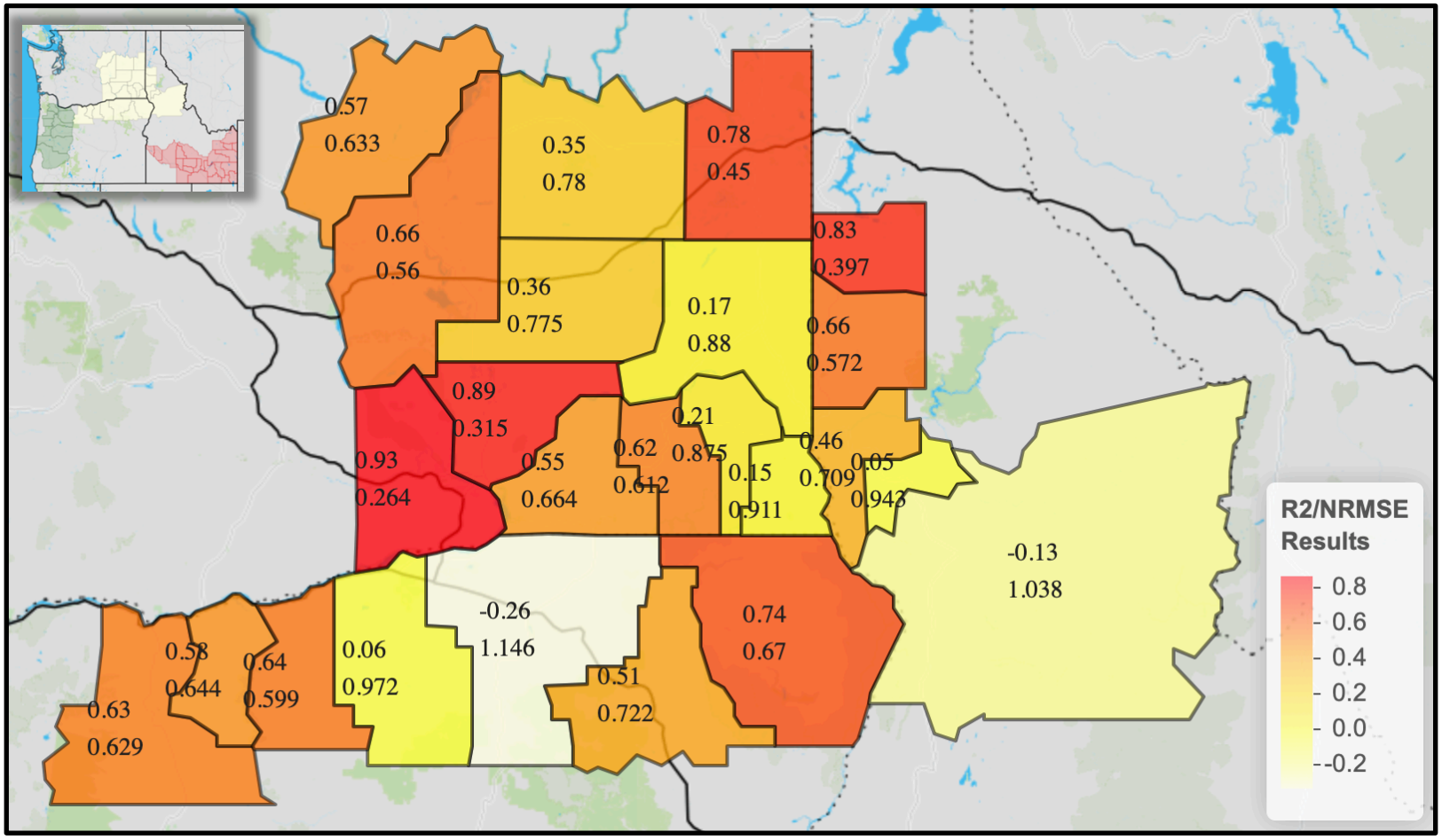


Figure 2.23. R^2 values (top value) for county wheat/drought randome forest model outputs, along with normalized RMSE (bottom value). Model performance is better in counties with less extreme loss variability.

CHAPTER 3: DEVELOPMENT OF A REPRODUCIBLE SCIENCE FRAMEWORK TO EXAMINE INLAND PACIFIC NORTHWEST (IPNW) AGRICULTURAL INSURANCE LOSS IN RELATIONSHIP TO CLIMATE

3.1. Introduction

Reproducibility is a critical component to scientific research. Over the last twenty years, the ability to generate more data, more rapidly, and with a greater variety, have led to an increased focus on data science methods for scientific evaluation, organization, and analysis (Zikopoulos et al., 2013). This expansion in data generation has resulted in more collaborative and transdisciplinary efforts, which has not only made reproducible science even more difficult (Blow, 2014) but has introduced questions about how reproducibility is defined. For example, language-specific variations can confuse the meaning of reproducible science (e.g. consider the words "replicable", "repeatable," or "generalizable"). Even the act of defining reproducibility across disciplines is onerous: in some instances, reproducibility is the test of a scientific hypothesis or experiment: if an experiment is repeated, the same resultant output is produced (Goodman et al., 2016). In medical and health fields, repeatability is generally described as a technique which measures the variation in measurements taken by a single instrument or person under the same conditions, while reproducibility measures whether an entire study or experiment can be reproduced in its entirety (Bartlett & Frost 2008; National Institute of Standards and Technology [NIST], 2007). Other disciplines, such as computational sciences, define reproducibility as a deterministic outcome, so that the re-running of an experiment should produce identical modeling results, which may include hyperparameter sampling decisions, or the random sampling nature of a training/testing model (Stodden et al., 2016). Furthermore, the concepts

of “open access” or “open science” complicate the discussion. While reproducibility, in a broad context, focuses on the ability to duplicate a scientific research study or experiment for the purposes of evaluation and replicability, open access and open science expand the discussion to include the ability to expose scientific processes to a universal audience, without the limitations of proprietary publication or repository access. Baker (2016) conducted a survey of over 1500 scientists, across a variety of scientific disciplines, specifically on this topic of reproducibility. The results were startling: more than 70% of those surveyed indicated that they had tried, and failed, to reproduce other scientists’ experiments, and over 50% failed to reproduce their own experiments. While the majority of respondents acknowledged that reproducibility is a significant crisis, less than 31% indicated that the failure to reproduce meant that a publications’ results were incorrect.

Popper (1992) notes that “communicating a scientific result requires enumerating, recording and reporting those things that cannot with advantage be omitted.” Under this premise, sound scientific research and analysis should effectively describe the methods and analytic steps required to support underlying hypotheses and subsequent results. In addition, Stark (2018) has proposed that the concept of “preproducibility” is a prerequisite for reproducibility. Preproducibility occurs if “an experiment or analysis has been described in adequate detail for others to undertake it”. Therefore, effective detail and information must be provided to document and produce analyses before such outcomes can be replicated, reproduced, or repeated. Putting this approach into practice, Johnson et al. (2017) attempted to examine the rates of reproducibility across 100 psychological science studies: 97% of the original studies showed significant results, while only 36% of those same studies, when reproduced, showed significance.

By extending the preproducibility/reproducibility process and introducing cross-disciplinary, interdisciplinary and transdisciplinary research, we create an added level of complexity that is extremely challenging, particularly with regards to data intensive analyses. While interdisciplinarity (Fairbairn & Fulton, 2000; Augsburg, 2005; Davies & Devlin, 2007; Pennington et al., 2016) addresses overlapping methods and ontologies between disciplines while persisting research separation, transdisciplinarity is a much more integrative effort that may combine methods, data, systems, and processes in a holistic way which results in a merged, unique research structure (Nicolescu, 2008). Transdisciplinary efforts not only relate to data, systems, and methods, but to people, and the ways in which they think about their own research (Gray, 2008). Stokols et al. (2013) indicate that transdisciplinarity "entails not only the integration of approaches but also the creation of fundamentally new conceptual frameworks, hypotheses, and research strategies that synthesize diverse approaches and ultimately extend beyond them to transcend preexisting disciplinary boundaries." As such, this form of collaborative science can be extremely valuable in terms of tackling complex problems, particularly those that span across multiple disciplines or that require unique, multi-pronged hypotheses to address the questions at hand. While transdisciplinarity is a complementary extension of traditional multidisciplinarity, Nicolescu (2008) further notes that it is "nevertheless radically distinct from multidisciplinarity and interdisciplinarity because of its goal, the understanding of the present world, which cannot be accomplished in the framework of disciplinary research". In this context, scientific transdisciplinarity is extremely beneficial, particularly in current scientific collaborative environments (e.g. climate change, biogenomics, astrophysics). While transdisciplinary research provides clear benefits to addressing multi-faceted research questions, it simultaneously is difficult to

successfully implement, and even more challenging in terms of its output preproducibility/reproducibility. Such research efforts have been popular in the several applied scientific areas such as bioinformatics and climate (Shi et al., 2008). Of additional note are the aspects of reliance upon data intensive processes in both example areas, as well as the importance of modular, reproducible scientific frameworks (Devezer et al., 2019). If we focus in on the importance of reproducibility and its relationships to transdisciplinary, data intensive research, we can identify numerous challenges, including:

1. The act of initiating transdisciplinary, reproducible research, given the interpersonal and discipline specific variations in terminology, research methods, data management, and ontologies.
2. Concerns around potential intellectual property rights, as well as more informal concerns around the competitive nature of scientific research and exposing work efforts that may be beneficial for other scientists who may be competing for similar funding.
3. The ability of researchers to quantify the lifecycle of their research in a timely fashion that aligns publications with output results, particularly given data-intensive research efforts.
4. Quantifying the methods and data in such a way as to easily expose reproducibility elements.
5. Issues of scaling in the number of researchers engaged, as singular efforts or smaller teams may have less resources to formulate systems and architectures which facilitate reproducible elements; similarly, larger, transdisciplinary teams may also experience

challenges in remote collaboration, communications, and effectively integrating research objectives.

6. Access to high performance computing (HPC) services and the barriers to engaging such services if research teams have little expertise in computational processes.
7. Issues regarding the long-term viability of reproducibility, given the need for data and code persistence and the cost of maintenance, particularly with regards to extremely large datasets. Stodden et al. (2010) did a survey of the machine learning community regarding the barriers to data and code sharing in the computational sciences: the most important issue raised was the time to document and clean up data (77%) and code (54%). Additionally, Vines et al. (2014) note that the data availability of research publications declines considerably with age.

These challenges can place heavy burdens on research teams, funding agencies and academic institutions, as well as journals, all of which have varying responsibilities to ensure reproducible science is achievable (Figure 3.1). In total, the lack of focus in these areas may hinder the initial expansion of such integrative research, as well as reducing the importance of reproducibility in general. The overall implications of such roadblocks are to discourage the type of progressive scientific discovery that is necessary for complex challenges facing our world. Counterbalancing these challenges, however, are numerous advantages that provide added incentive to researchers and their institutions, including: (a) the overall value of integrated scientific research in exploring complex systems, such as biological, environmental, socio-economic, or combinatory aspects (Cooke et al., 2015); (b) uncovering scientific issues or questions that were previously unexplored or unknown (Stokols et al., 2013); (c) establishing collaborative relationships across domains that may persist for

engagement in other projects or aspects of research; (d) enhancing individual domain research directions by introducing alternative methods or techniques that were heretofore not known or utilized (Uzzi et al., 2013); (e) enhancing funding streams for research efforts which may enhance systems, processes, or data environments, thereby improving future research capabilities (Wuchty et al., 2007); and (f) having an overall positive value on science impact (Wuchty et al., 2007; Uzzi et al., 2013). From a more refined perspective, there are specific, tangible outcomes that indicate value from transdisciplinary research efforts, including: group publications are more highly cited than publications by individuals, which serves as an indicator of impact (Wuchty et al., 2007). Compared with solo authors, teams and groups across disciplines were also more likely to develop research that organized unique ideas into high-impact publications (Cooke et al., 2015). Stipelman et al. (2014) performed a meta-analysis of publications produced by disciplinary vs. transdisciplinary research centers over time, concluding that transdisciplinary team science approaches tend to broaden the reach of research findings more so than traditional disciplinary efforts.

Given the aforementioned challenges and added value regarding reproducible science and its relationships to data intensive research, the focus of this work was to quantify and prioritize the methods and processes that are necessary for reproducible, data intensive research, with a particular emphasis on transdisciplinary efforts. We lay out a process-oriented framework for such a structure, using two complementary disciplines for our case scenario: agricultural systems and climatology. Our focused goal is to document and construct an effective data intensive research framework using these complementary disciplines, providing actual systems, data, and code that could assist other researchers in furthering their scientific

endeavors. As part of our methodology, we address specific challenges and pitfalls to provide solutions, spanning both technological as well as collaborative team building facets.

3.2. Methods

Our methodological approach is divided into two components: (a) description of a framework which supports and facilitates reproducible science efforts, and (b) a case study data analysis example that applies the framework, using agricultural and climatic data as a foundation. This framework development and case study example were funded as part of the Climate Impacts Research Consortium (http://pnwcirc.org), a science-to-action team funded by the National Oceanic and Atmospheric Administration (NOAA). CIRC is also one of NOAA's Regional Integrated Sciences and Assessments (RISA) teams under award # NA15OAR4310145. Of particular note are the considerable efforts to establish similar data intensive frameworks and methodologies for reproducibility, including Flathers and Gessler (2018), Donoho et al., (2007) and Dumontier and Wesley (2018). Our proposed approach builds upon this research, extending the aspects of data input, processing, analysis, and storage, and includes facets of team collaboration, code storage, data management, project management, and code version control. Our premise is that successful reproducible science includes more than just data intensive analysis components, and includes aspects of scientific collaboration, team development and leadership.

3.2.1. Reproducible scientific framework (RSF) components

The framework methodology identifies the most important facets of transdisciplinary, reproducible science, and quantifies such a framework in a proposed set of systems, data,

code, analytics, and management processes. As part of our methodology, we lay out each of the individual framework sections below, which are also depicted in Figure 3.2. We then describe our prototype case study, the relationships to the framework, the challenges and limitations encountered, and the implications to other reproducible science projects.

3.2.1.1. Data management and the FAIR principles

At the core of the ability to develop sound, modular efforts for reproducible, open science efforts, is the ability to organize, manage, and persist data (McKiernan et al., 2016). To a greater extent, such data must conform to a particular structure to allow for eventual reproducibility. Wilkinson et al. (2016) outline the FAIR principles (findability, accessibility, interoperability, reusability), a recent set of data management concepts that, at a high level, set a structure for reproducibility and open science efforts which ultimately support the tenet of scientific transparency. Critical to enabling the FAIR principles is the ability to structure and organize data in a way that permits transdisciplinary research interaction and reproducibility. These efforts are not trivial and can require knowledge in the areas of system and network administration (Bergstra & Burgess, 2008), metadata specifications and ontologies (Giles, 2011; Mayernik et al., 2016), database architectural development (Wandell et al., 2015) as well as integrative software development (Hermann & Del Balso, 2017). Each of these areas, on their own, have considerable challenges and complexities; to assemble them in a cohesive fashion can create difficult hurdles which can discourage the type of integrative research efforts described previously. Numerous data management-specific systems exist: SciTran (http://scitran.github.io), developed by Stanford University's VISTA lab (Wandell et al., 2015), is a recent effort to facilitate scientific

transparency and reproducibility through effective data management approaches, which is instantiated as part of a RESTful (representational state transfer) application programming interface (API). SciTran has recently been migrated to a for-profit software package called FlyWheel (http://flywheel.io) and is used primarily in medical research data management efforts. Tableau (http://tableau.com) is another for-profit software system that is heavily used in commercial and research settings for data management, visualization and analysis purposes. From an open-source perspective, the Comprehensive Knowledge Archive Network (CKAN - http://ckan.org) is one of the most utilized data management platforms worldwide, used by a number of governmental and academic institutions for data storage, retrieval, and metadata documentation (Winn, 2013). CKAN is used by the United States federal government in numerous capacities, most notably as the engine for http://data.gov. It is also the framework used by the European Union's European Data Portal (https://www.europeandataportal.eu), Australia's Commonwealth Scientific and Industrial Research Organization (CSIRO), as well as the governments of Canada, Switzerland, and a large number of municipalities and cities (http://ckan.org).

Apache Solr (http://apache.org) is another data management platform that is based on Apache Lucene, a set of Java-based indexing and search libraries, which provide spellchecking, hit highlighting and advanced analysis/tokenization capabilities (Shahi, 2015). Apache Solr has numerous advantages: it is highly scalable and fault tolerant, is enterprise ready, can accommodate a full text search, is extensible, and has relatively easy configuration and administration. For the purposes of our case study example, we implemented Solr on Linux Red Hat Enterprise Server 6.0, for the storage of agricultural datasets from the National Agricultural Statistics Service (NASS) and the USDA risk management agency

(RMA). Given Solr's RESTful access over HTTP, it allows for easy incorporation of metadata search requests into other software deployments.

Operationally supporting the FAIR principles and data management systems are a large grouping of research models that attempt to standardize data management efforts across differing domains. An established approach has been to construct service-oriented architectures (SOA), a concept from computer sciences that describes building modular, loosely coupled software systems which expose content using APIs, integrating disparate domains for the purpose of diverse research and analysis (Papazoglou & Van Den Heuvel, 2006; Flathers et al., 2017). By utilizing a SOA, transdisciplinary research efforts can associate defined database and modeling systems in a way that allows for autonomy and flexibility, while preserving legacy environments that may be difficult to migrate or transform (Pessoa et al., 2008). In total, data management systems which enable machine-ready requests, and facilitate distributed data coupling with other analysis and modeling efforts through a SOA, are a product of the need to reproduce scientific processes without the burden of manual data integration, searching, or access. As such, they are a critical component of not only large research efforts, but smaller teams that nonetheless utilize large amounts of data and processing to derive results.

3.2.1.2. Dynamic data requests

Enabling reproducibility and iterative analysis methods, particularly given large data that may be spatially and temporally variable, often times requires the efficient ability to re-run analyses using differing data transformations (which may entail differing variables, geographies, and time scales). Using climatic information as an example, we note the huge

array of data that exist for atmospheric, oceanic, surficial and groundwater, as well as paleo-climatic data, that are often structured using existing specifications and formatting (Table 3.1). In addition, there has been extensive research into the processes of downscaling and structuring climatic data for gridded analysis and use (Daly, 2006; Thorton et al., 1997; Abatzoglou, 2013). In the climatological disciplines, NetCDF (Network Common Data Form) datasets are the standard to represent such multi-dimensional array data, that may span many variables across differing time frames, and that are typically in a gridded spatial structure. Developed by the University Consortium for Atmospheric Research (UCAR), and based on the National Aeronautical Space Administration's (NASA) CDF model, NetCDF represent a set of common libraries which allow for machine-independent access (Treinish & Gough, 1987) and is the de facto standard for representing such data by academia, government, and corporate entities worldwide (Harnett & Raw 2008). To a greater extent, such climate and forecast convention (CF) formatting is commonly used for global climate modeling (GCM), as well as coupled general circulation modeling efforts which forecast /simulate meteorological outcomes into the future (Eaton et al., 2011). Several mechanisms exist for accessing and transforming multi-dimensional gridded datasets, including distributed data access server platforms, such as Thematic Real-Time Distributed Data Services (THREDDS – https://www.unidata.ucar.edu/software/tds), Hyrax (https://www.opendap.org/software/hyrax-data-server), or ERDDAP (https://www.ncei.noaa.gov/erddap/index.html). THREDDS, Hyrax, and ERDDAP are web servers which organize metadata and content for scientific datasets (typically gridded), using a variety of remote data access protocols, such as the open-source project for a network data access protocol (OPeNDAP), the Open Geospatial Consortium (OGC), Web Coverage

Services (WCS), and HTTP. OPeNDAP in particular is a useful protocol, which enables the use of data from a remote server, without the need to download data files, and includes metadata inspection, sub-setting, slicing, and aggregation of data (Cornillon et al., 2003). While OPeNDAP allows for representational state transfer (REST) requests via a web browser, these same data queries can be implemented directly from other scientific software, such as Python, Java, Matlab, and R.

THREDDS specifically provides a gridded array model for NetCDF data storage and access, extending access by permitting dynamic aggregation and subsetting of gridded data files using RESTful API requests. For example: a user may have a particular need to access a specific number of variables, for a refined geographic location across a limited time frame. Instead of having to download large amounts of data, which may include extraneous information that is not needed, they can formulate a request (utilizing OPeNDAP that uses the THREDDS RESTful API via the NetCDF subset service, or NCSS), to acquire only the needed data, aggregated or subset in a fashion that fits their purposes. ERDDAP, a server architecture developed by NOAA, is also useful for quickly building RESTful URLs which generate outputs in a variety of structures. Similar to THREDDS, such dynamic URLs can be embedded into software code, and/or manipulated to change or refine input variables. These protocols and platforms facilitate the ability of researchers to not only modularly structure analyses, but to easily modify such analyses for alternative hypotheses, and publish the base code for reproducibility purposes, without providing the mass of data utilized as part of that process.

3.2.1.3. Numerical model integration

An important advancement in reproducible scientific efforts has been the integration and re-use of models from different disciplines to simulate complex environmental systems, such as the Community Surface Dynamics Modeling System (CSDMS) (Peckam et al., 2016). Other related model integration efforts include the Earth System Modeling Framework (ESMF) and the Open Modeling Interface (OpenMI) (Gregersen et al., 2007). These frameworks are a considerable advance in the ability of the geosciences discipline to couple various numerical models together using component-based programming, across differing programming languages and operating systems, thru the use of open source software standards. Such modeling integration efforts are a critical piece to the future of reproducible and replicable science, by loosely connecting established modeling outputs together. There are numerous benefits that include performance, ease of maintenance, ease of use, flexibility, portability, stability, encapsulation, and software longevity (Overeem et al., 2013; Peckam et al., 2016).

3.2.1.4. Dynamic data analytics

Data analyses are typically the final output of any hypothesis-focused and/or data-driven research. Such outputs may have a statistical emphasis and can include time series and/or geospatial relationships, as well having associations with differing discipline-specific methodologies (Daley, 2006). A key conclusion, from a broad perspective, is that with more data inputs, that span multiple disciplinary areas, reproducibility and method standardization become more complex (Mesirov, 2010). For example: agriculturally-focused research into understanding agricultural insurance loss and the relationships to climate may involve (a) multiple cropping systems, (b) insurance loss data collected over multiple counties and/or

time points, (c) additional agricultural variables, such as management practices, biophysical and/or genetic components of the particular cropping systems, as well as (d) meteorological/climatic data, across multiple variables, that similarly may have variations temporally and spatially (Seamon et al., 2019b). In this basic example, we have potentially introduced a multitude of data sources and methods that make the core analysis portions fairly complex for allowing reproducible results. A particular researcher may be able to establish a singular set of procedures that effectively allow for the preproducibility of analyses, and to document the procedures within a publication, but is such documentation sufficient for reproducible results? Likely not. Constructing the output results in the form of a dynamic set of analytic tools, which allow for the iterative running/re-running of processes, as well as exposing the standardized code for such analytics, provides a more effective foundation for facilitating reproducibility (Guru et al., 2016) Extending this premise, the assumption is that an external researcher may not have the full depth of knowledge of the preproducible results, and therefore will likely require a roadmap/workflow of the analytics, with accompanying data and code, in order to align reproducible efforts with the steps of the original analysis.

3.2.1.5. Modular code development

A critical facet to most aspects of transdisciplinary scientific research is the ability to collaborate on programming code. To a greater extent, standards that allow for the modular construction of such code, in a way that allows for the ease of re-use, reproduction, or transformation of modeling outputs, is an important part of reproducible science. There have been a number of developments to facilitate such collaboration. For example, The

ability to collaborate and reproduce code/modeling outputs through the use of dynamic research notebooks has been a paradigm shift in the quantitative sciences (Kluyver et al., 2016). One of the most popular efforts in this area is the Interactive Python (IPython) research notebook development (Perez, 2007). In 2014, the notebook portion of IPython was separated from its kernel and shell, and re-named Project Jupyter. Jupyter notebooks support a variety of programming languages (R, Ruby, Python), and leverage a web service interface to allow programming collaboration in a dynamically run window via HTTP. Up until the development of IPython/Jupyter, teams were required to utilize a code repository/version control framework, such as Apache Subversion (http://subversion.apache.org, 2019), or more recently, cloud-based code repository frameworks such as Github (http://www.github.com). Dynamic research notebooks do not replace such code repository/version control approaches, but rather complement their usage by allowing for real-time collaboration and the dynamic presentation of code output, with accompanying visualizations and statistical outputs. To elaborate on this complementary nature, code repository systems such as Github allow for the storage and version control of .ipynb or .Rmd files, which are the outputs of research notebook systems such as Jupyter or R/Rstudio.

3.2.1.6. Referencing data/publications using Digital Object Identifiers (DOIs)

A digital object identifier uniquely denotes a dataset or publication, by creating an electronic object which contains metadata used to reference the materials (International DOI Foundation, 2012). DOIs are structured in such a way as to allow for machine ingestion, which enables the intelligent searching and filtering based on keywords, subject topics, data types, or other metadata content. DOI usage has expanded considerably into data, code, and

other related research outputs, particularly given the growth of data repositories for permanent data storage. Some key facets of DOIs include:

1. Discoverability. By attaching a unique identifier that can contain metadata, datasets can not only be associated with related publications, but can be searched and queried via research engines such as Google, ORCID, dryad, CrossRef, or datecite. Such discoverability only extends that capacity of research efforts to be replicated or extended.
2. Availability and/or persistence. A core premise of the DOI architecture is that only datasets which reside in a relatively permanent location can be assigned an identifier (Kahn & Wilensky, 2006). By hosting datasets in a permanent home, and assigning a DOI, the likelihood of reuse increases.
3. Scientific impact. The combined aspects of discoverability and persistence facilitate/increase the overall scientific impact of a publication and its related data. By establishing a level of stability in terms of the full understanding of a research effort, the potential for discoveries to effect related or peripheral research is magnified (Michener, 2015).

3.2.1.7. Workflow construction

Workflow templates are a commonly used technique to outline data intensive scientific processes and are a useful tool in term of organizing and structuring methods which can then be used as to facilitate replication/reproducibility (Deelman et al., 2008). As teams work to develop integrated research processes, it may be easier to develop these efforts individually, with little interaction between subsets of the overall team, and cobbling together the results

of such analysis at a later stage. However, with more complex research questions, this form of process organization may be limiting: similar to other transdisciplinary functions, the task of integrating such processes across disciplines may be difficult and time-consuming. As a whole, developing an effective workflow strategy that encapsulates the spectrum of research needs will benefit long-term analysis capabilities, and extend the prospect of reproducibility or added value in derivative research efforts.

3.2.1.8. Collaborative research communications

Often times the aspect of qualitative research communications are underestimated in terms of the overall value to reproducible, transdisciplinary science. With an essential focus on data assembly, processing, modeling outputs, and permanence of storage, the methods and approaches for effective collaboration are usually ranked low in terms of importance (Russell et al., 2008). The ability for scientists of differing disciplines (and sometimes in differing locations) to communicate effectively, and in turn, facilitate collaboration at a level that spurs scientific discovery, is not trivial. Traditional scientific research has, for many years, been framed in the context of the "silo" mentality (Gray, 2008). Researchers are, at the most basic level, solely responsible for their own methods, analyses, and approaches to organize and structure their research program, and such approaches may vary dramatically based on expertise, technology capabilities, and discipline (Martín-Sempere et al., 2008). Adding to this complexity are data-intensive research efforts, which may occupy tremendous amounts of time and energy in order to merely get to the point of being able to run a model or perform analyses. Under this paradigm, a researcher is forced to prioritize needs, which includes the time needed to understand and bond with fellow collaborators. Such review is continuously

being performed by focused researchers. Is there value in attending this collaborative meeting? What deadlines exist for me in terms of other research commitments? Am I honestly going to derive value from this collaborative communication? When research teams undertake difficult transdisciplinary efforts, the focus is usually on the technical, operational, scientific, and data-specific requirements that are needed to explore the proposed scientific hypotheses. Yet in many ways, the ability of researchers to acknowledge the value in collaborative communications, and prioritizing such efforts, is critical to succeeding in such transformative research efforts (Bennett & Gadlin, 2012).

There are several approaches which attempt to bolster such collaborative scientific integration, including work done by Eigenbrode et al. (2007), who developed the Toolbox Dialogue Initiative (http://tdi.msu.edu/research/), an effort to facilitate improved communications for team based collaborative science efforts. The toolbox effort is founded in a dialogue-based workshop, where researchers discuss their beliefs and values in relationships to scientific collaboration and team interaction. Using this tested dialogue approach, interdisciplinary and transdisciplinary teams can enhance their abilities to achieve project success (Morse, 2013). Such team engagement is an essential aspect of integrated science that involves expertise across a spectrum of disciplines, and particularly so when such research involves data-intensive science.

3.2.1.9. Content integration

Framing all of the previous components, in a manner that encourages their use and examination, has typically resulted in some form of a dynamic website. However, depending upon the nature of a research team, and their interaction with external partners and/or

stakeholders, such a communication mechanism takes on multiple roles. Not only does it communicate and facilitate research efforts by allowing team members to access systems and related instructional materials for the use and operation of those systems, but it may also serve to communicate scientific outcomes to differing audiences in differing ways. While the development of such content integration is likely seen as trivial in relation to other aspects of reproducibility, the failure to effectively devote time and energy to such tasks can hinder overall team and stakeholder interaction (Böcher & Krott, 2014). Aspects of such content integration includes scientific communications writing, user interface design, as well as scientific tool assessment and evaluation (Bagstad et al., 2013).

3.2.1.10. Scientific project management

Structured scientific project management to implement and iteratively improve all of the previously mentioned components, is another overlooked factor that ensures reproducibility. Many research projects which have worthy goals and objectives fall to the wayside given poor planning, no real structured model for resilience and longevity, or an effective data management plan that incorporates maintenance and support beyond the life of the funding. Similarly, few academic institutions have organizational models that recognize the need to address these long-term aspects of resiliency, from a resource and financial perspective. If a funding agency makes a considerable investment in a research project, how will the long-term implications of such a project affect access to the underlying data, code, modeling, and analytics that may produce resultant publications? Is there not an added value in research efforts that spur scientific discovery, based on the foundational research outputs that support initial publication development? The general consensus amongst wide swaths of research

communities affirms that there is considerable value in making such information available, as it enables pyramid-like research generation (Mesirov, 2010). Ensuring that such long-term planning and research resiliency exists requires effective project management efforts, that are, in and of themselves, a research component of many transdisciplinary efforts. Engaging researchers whose core focus is in such areas would be a wise consideration in diverse, large project teams, particularly those that involve intensive data science-driven goals. Project management is an established discipline with an extensive body of research, across a variety of scientific and non-scientific areas (Brocke & Lippe, 2015).

There are numerous project management methodologies, but the Agile methodical framework is one which is used regularly in corporate and governmental software development organizations (Cockburn & Highsmith, 2001). Agile emphasizes iterative, spiral-like development processes, with self-organizing teams providing the majority of emphasis on tasks and timelines (vs. a formal project manager who may be more isolated from the detailed work effort). While academic research settings are not corporate environments, they would be well served to take the best-of-breed portions of such project management efforts, balancing that with academic creativity and scientific discovery, which is not always set to a particular time frame. Nonetheless, more of a focus on structured project frameworks may assist transdisciplinary teams in their organizational data-intensive research efforts, and thus, improve the abilities to construct reproducible outcomes.

3.2.2. Case Study Example: Climate and Agriculture

Our case study data analysis example focuses on agricultural insurance loss analysis and the relationships to climate in the inland Pacific Northwest US (iPNW). While our overall

hypothesis was that climatic outcomes have a temporal and spatial predictive power in terms of commodity-specific insurance loss, the purpose of the case study was not merely to provide the quantitative results of the analysis, but also to exemplify how the proposed framework could be used to facilitate reproducibility in a particular analysis scenario. The detailed analysis results of this case study can be found in two additional publications, Seamon et al. (2019a) and Seamon et al. (2019b). We used two data sources for this example analysis:

1. Agricultural crop insurance. Crop insurance claims from 2001 to 2015 were acquired from the United States Department of Agriculture's (USDA) risk management agency (RMA). The USDA's RMA provides an extensive archive of crop insurance data, at an individual claim level, by commodity, county, and year, along with specific information about the cause of the damage and the total loss ($) of the particular claim. In the Pacific Northwest alone (Oregon, Idaho, Washington), over 20,000 agricultural insurance claims were filed across 35 differing commodities rom 2001 to 2015 (Seamon et al., 2019a). Data sources for insurance loss are accessible in Appendix D.

2. Gridded climatological data. In addition, we utilized daily gridded climate data at a 1/24th degree spatial resolution (~4km/pixel) (Abatzoglou, 2013), acquired from the University of Idaho's (UI) THREDDS server, which is hosted by the Northwest Knowledge Network (NKN – http://northwestknowledge.net). NKN provides research data management and computing support for UI researchers and their regional, national, and international collaborators. Data sources for climatology are

accessible in Appendix D. Using these two datasets, we constructed a step-by-step analysis workflow, which is depicted in Figure 3.3. Our case study workflow is described below in terms of analysis steps. The overall reproducible science framework provides support for this analysis model, with each framework component having an associated case study parallel (Table 3.1).

3.2.2.1. Agricultural data acquisition and organization.

In our initial step, we downloaded agricultural insurance loss files for the conterminous United States from 1989 to 2015. Files were available in a comma separated format from the USDA web site (http://usda.gov/rma). Insurance loss data were combined in R and aggregated by several factors (year, county, damage cause, and cropping system), which created several transformed datasets. The resultant data and code functions for step 1 were uploaded to a central GitHub repository, and are noted in Appendix A. The outputs of this aggregation allowed us to explore the totality of agricultural insurance loss for the study area by commodity, county, year, and by the cause of damage (e.g. drought, heat, excessive moisture).

3.2.2.2. Data transformation and exploratory data analysis.

In our second step, data mining processes were performed on our transformed agricultural insurance loss data, in order to evaluate issues of missing data and spatial/temporal variability by crop type. For example, principal components analyses (PCA) was performed to evaluate factorial relationships, which resulted in a reduction of our spatial and temporal extent, as well as focusing on wheat insurance loss due to drought for the inland Pacific

Northwest (iPNW) region (Figure 3.3). Finally, the insurance loss data for this refined area were prepared for associations with climate data. The resultant data and code functions for step 2 were similarly uploaded to a central GitHub repository, with procedural steps described in appendices A and B. A fully detailed review of the analysis results from steps 1 and 2 can also be found in Seamon et al (2019a).

3.2.2.3. Climate data acquisition and dataset combination.

In our third step, we acquired and transformed climate data (precipitation, maximum daily temperature, potential evapotranspiration, and the Palmer Drought Severity Index (PDSI)) from the University of Idaho's THREDDS server, using a combination of RESTful API requests, as well as NC operator statements (Zender, 2006). Given that the available data was accessed at a gridded and daily timestep, we needed to aggregate to a monthly basis, by county. Finally, the code functions which performed this spatial and temporal aggregations were used to generate a range of monthly combinations, which were then used to evaluate optimum correlational relationships to insurance loss for wheat, due to drought. The code and analysis portions of this process were documented in RMarkdown and are referenced in Appendix C.

3.2.2.4. Predictive modeling using analytic dashboard development

In our final step, we used the data outputs generated for our analysis to construct a set of dynamic data analysis and predictive modeling dashboards within R, exposed as web applications using R's web server, called Shiny (https://www.rstudio.com/products/shiny/shiny-server/). Shiny server works in conjunction

with R, to expose R programming via a dynamic web interface, allowing users to explore data analysis and modeling outputs on their own. Several specific dashboards were developed that address a range of analytic capabilities, with a focus on (a) exploratory data analysis, and (b) predictive modeling. A list of dashboards and their online access URLs are noted in appendices F and G. Our dashboard development had several advantages: it allowed us to use a modular approach to document our analysis processes, for iterative data examination and visualization. With minimal alterations, we were able to clone multiple versions of dashboards to present varied versions of predictive modeling outputs (gradient boosted regression, random forest, neural networking), as well as a multiple exploratory data analysis outputs. Additionally, by using a web-based application interface, we were able to expose the analysis capabilities to a wide audience, for review and use. Such approaches, combined with research notebook outputs (Jupyter, Rmarkdown), allow for a full spectrum understanding of research approaches to problem sets, which promotes reproducibility in multiple forms.

3.3. Discussion and Conclusions

Our case study analysis identified several challenges regarding reproducible science in transdisciplinary, data-intensive research efforts, spanning technological, operational, and communications/collaboration areas. In addition, the challenges raised here have a combinatory effect, particularly in big data/data intensive research areas that overlap disciplinary methodologies. As such, the proposed framework provides a frame of reference for future research teams who see the value in outputs which are reproducible, for themselves

as well as for additional researchers working on related problems. The results of our case study analysis which uses the described framework indicated several takeaways:

Processing efficiency and challenges. With larger datasets, the ability to access and manipulate such information remotely (e.g. use of API requests) can be difficult depending upon server capabilities and network speeds. As part of our case study analysis, we needed to perform some data transformation using NCO (NetCDF Operators) with local access to climate data. NCO are open source command line functions which work with NetCDF, HDF, or DAP files, and assist in analyzing gridded or unstructured scientific information (Xu et al., 2019). For our purposes, we integrated these transformations into Linux/bash scripts, which are additionally available online, as noted in Appendix D. However, a key takeaway from this process step is that API data access is not necessarily a complete solution in all cases when large, complex data transformations are needed. These issues continue to challenge researchers in a variety of data-intensive scientific disciplines.

Data persistence and embargo methods. Aspects of data persistence over the long term is a critical component to long-term reproducibility and viability (Michener, 2015). As data modifications are made over time, issues of reproducibility and replication may be affected, particularly when time-series content is involved (e.g. climatological data generation). Embargoing is a technique for data inclusion with a delay time period before access is enabled. Such embargoing is important if the publisher believes that said data may change or be altered. However, embargoing does not necessarily address regular, on-going changes to datasets that are valid at each time point. In the instance of parent-child data relationships, not only is the core data need to be preserved, but the metadata associated with such

information (instrumentation usage, calibration methods, bias correction processes) is important in order to enable reproducibility (Hampton & Parker 2011; Bowser et al., 1994).

Scaling. Issues of temporal and spatial scaling are a continued source of concern in large data analyses, and present on-going issues in terms of reproducible science. Flathers and Gessler (2018) note these issues in their efforts to combine and analyze datasets at differing spatial resolutions, particularly regarding issues of data variance smoothing. In our case study analysis, there were issues of scaling both spatially and temporally: with agricultural data at a county/monthly scale, we were forced to aggregate our daily/4km per pixel climate data to a coarser spatial and temporal resolution, in order to standardize our analyses. While not ideal, this form of accommodation is not uncommon, given the limitations of available data, and the scaling structure that may be imposed given confidentiality or other organizational constraints (National Research Council, 2001).

Time commitments and efforts for reproducibility and modularity. Modular, reproducible science takes time and energy, and is very often underestimated. The amount of effort to construct an effective data management strategy, to agree upon scientific nomenclature and ontologies, as well as continuing to keep long-term processes in mind when code is developed, are time consuming and may be neglected in deference to expediting the research outcomes. As with all research programs, a level of balancing must take place in terms of available resources and time commitments, while ensuring that the methodological approaches will support reproducibility in an appropriate fashion.

Commitment to modularity in code development will pay off when reanalysis is performed. As part of the case study development, a considerable challenge was to design functions and

related code with a mindset of modularity. For example, we tried to use discrete functions for portions of analysis, data aggregation, climate analysis, and modeling, that were self-contained: Under this approach, the functions could be used separately for differing needs, or to reproduce portions of the work for validation and replication. However, such work took a considerable amount of time, and required a focused effort to plan, document, and organize the code and related data in such a way as it could be understood by someone outside of the project. While this type of work is not necessary to construct the scientific analyses or models, it is important to stress these tasks as part of a project that wants to encourage reproducibility.

Willingness to use alternative methods if the need arises. A fairly common adage that has been used in many capacities, but is applied frequently in the statistical modeling and machine learning community recently, is the following: "if the only tool you have is a hammer, everything looks as if it were a nail" (Kaplan, 1964; Kolby, 1963; Maslow, 1965). In terms of modeling, a data scientist who is familiar with a particular form of analysis may be more inclined to use this approach frequently, even if such an approach is not necessarily warranted. From a data analysis and applied data science perspective, R and python are two approaches to scientific programming, which require an investment of time for applied use. A takeaway from our case study work was that both of these programming environments have value in differing capacities and may complement each other in certain situations (factoring in resource availability and allotted time).

Ability to apply modular techniques to unique analytical needs. Most analysis and research efforts have very particular goals and data needs. Given these constraints, one might

envision that each project effort would require a unique development process, with little commonality to other project efforts. Yet by applying some of the noted standardized approaches (use of dynamic data access requests, use of data/code repositories and version control, code collaboration, agile project management), a basic research framework is established which makes reproducibility more achievable. If these approaches are combined with other institutional mechanisms (required use of DOIs, publishing of data in conjunction with publications, funding of workflow provenance and reproducibility standards), then a wide spectrum of differing project efforts can considerably improve the likelihood of reproducibility.

3.4. References

Apache Solr. (2019). Retrieved from http://apache.org/solr

Atkins, D. E. (2012). Data Stewardship in the Age of Big Data. ERCIM News, (89).

Augsburg, T. (2005). Becoming Interdisciplinary: An Introduction to Interdisciplinary Studies, 3rd Ed.

Bartlett, J. W., & Frost, C. (2008). Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. Ultrasound in Obstetrics and Gynecology, 31(4), 466–475. https://doi.org/10.1002/uog.5256

Bagstad, K. J., Semmens, D. J., Waage, S., & Winthrop, R. (2013). A comparative assessment of decision-support tools for ecosystem services quantification and valuation. Ecosystem Services, 5, 27–39. https://doi.org/10.1016/j.ecoser.2013.07.004

Baker, M. (2016). Is there a reproducibility crisis? Nature, 533, 3–5.

Bennett, M, & Gadlin, H. (2012). Collaboration and Team Science: From Theory to Practice. Journal of Investigative Medicine, 60(5), 768–775. https://doi.org/10.1038/jid.2014.371

Blow, N. S. (2014). From the Editor: A Simple Question of Reproducibility, 2144. https://doi.org/10.2144/000114117

Böcher, M., & Krott, M. (2014). The RIU model as an analytical framework for scientific knowledge transfer: the case of the "decision support system forest and climate change." Biodiversity and Conservation, 23(14), 3641–3656. https://doi.org/10.1007/s10531-014-0820-5

Brocke, J., & Lippe, S. (2015). Managing collaborative research projects: A synthesis of project management literature and directives for future research. International Journal of Project Management, 33(5), 1022–1039. https://doi.org/10.1016/j.ijproman.2015.02.001

Chervenak, A., Foster, I., Kesselman, C., Salisbury, C., & Tuecke, S. (2000). The data grid: Towards an architecture for the distributed management and analysis of large scientific datasets. Journal of Network and Computer Applications, 23(3), 187–200. https://doi.org/10.1006/jnca.2000.0110

Cockburn, A. and Highsmith, J. (2001). Agile Software Development: The People Factor. Computer, 34, 131-133.

Cooke, N. J., Hilton, M. L., & Sciences, S. (2015). Enhancing the Effectiveness of Team Science. Enhancing the Effectiveness of Team Science. https://doi.org/10.17226/19007

Cornillon, P., Gallagher, J., & Sgouros, T. (2006). OPeNDAP: Accessing data in a distributed, heterogeneous environment. Data Science Journal, 2(October), 164–174. https://doi.org/10.2481/dsj.2.164

Daly, C. (2006). Guidelines for assessing the suitability of spatial climate data sets. International Journal of Climatology, 26(6), 707–721. https://doi.org/10.1002/joc.1322

Davies, M. & Devlin, M. (2007). Interdisciplinary higher education: Implications for teaching and learning. Centre for the Study of Higher Education, the University of Melbourne.

Deelman, E., Gannon, D., Shields, M., & Taylor, I. (2009). Workflows and e-Science: An overview of workflow system features and capabilities. Future Generation Computer Systems, 25(5), 528–540. https://doi.org/10.1016/j.future.2008.06.012

Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in Scientific Data Infrastructure. In Proceedings of the 2013 International Conference on Collaboration Technologies and Systems, CTS 2013 (pp. 48–55). https://doi.org/10.1109/CTS.2013.6567203

Devezer B., Nardin L.G., Baumgaertner B., Buzbas E.O. (2019). Scientific discovery in a model centric framework: Reproducibility, innovation, and epistemic diversity. PLoS ONE 14(5): e0216125. https://doi.org/10.1371/journal.pone.0216125

Donoho, D., Stodden, V., Tsaig, Y. (2007). About Sparselab. Stanford University, Version 2.0. March 2007

Dumontier, M., & Wesley, K. (2018). Advancing discovery science with fair data stewardship: Findable, accessible, interoperable, reusable. Serials Librarian, 74(1–4), 39–48. https://doi.org/10.1080/0361526X.2018.1443651

Eaton, B., Gregory, J., Drach, B., Taylor, K., Hankin, S., Caron, J., … Juckes, M. (2011). NetCDF Climate and Forecast Metadata Conventions, v 1.6.

Eigenbrode, S. D., O'Rourke, M., Wulfhorst, J.D., Althoff, D., Goldberg, C.S., Merrill, K., . . . Bosque-Pérez, N.A. (2007). Employing philosophical dialogue in collaborative science. Bioscience 57(1):55-64. https://doi.org/10.1641/B570109

Fairbairn, B., Fulton, M.E. (2000). Interdisciplinarity and the Transformation of the University of Saskatchewan, Centre for the Study of Co-operatives, Saskatoon

Flathers, E., Kenyon, J., & Gessler, P. E. (2017). A service-based framework for the OAIS model for earth science data management. Earth Science Informatics, 10(3), 383–393. https://doi.org/10.1007/s12145-017-0297-3

Giles, J. R. A. (2011). Geoscience metadata: No pain, no gain. In A. K. Sinha, D. Arctur, I. Jackson, & L. C. Gundersen (Eds.), Societal Challenges and Geoinformatics (Vol. 482, p. 0). Geological Society of America. https://doi.org/10.1130/2011.2482(03)

Godlee, F. (2008). Open access to research. BMJ (Clinical Research Ed.), 337(7665), a1051. https://doi.org/10.1136/bmj.a1051

Goodman, S.N., Fanelli, D., & Ioannidis, John P.A. (2016). What does " research " mean ? Sci Transl Med, 8(341). https://doi.org/10.1108/CG-10-2012-0073

Gray, B. (2008). Enhancing Transdisciplinary Research Through Collaborative Leadership. American Journal of Preventive Medicine, 35(2 SUPPL.), 124–132. https://doi.org/10.1016/j.amepre.2008.03.037

Gray, J., Liu, D. T., Nieto-Santisteban, M., Szalay, A. S., Dewitt, D., & Heber, G. (2005). Scientific Data Management in the Coming Decade. Microsoft Research Technical Report MSR-TR-2005-10, (January).

Gregersen, J. B., Gijsbers, P. J. A., & Westen, S. J. P. (2007). OpenMI: Open modelling interface. Journal of Hydroinformatics, 9(3), 175–191. https://doi.org/10.2166/hydro.2007.023

Guru, S., Hanigan, I. C., Nguyen, H. A., Burns, E., Stein, J., Blanchard, W., … Clancy, T. (2016). Development of a cloud-based platform for reproducible science: A case study of an IUCN Red List of Ecosystems Assessment. Ecological Informatics, 36, 221–230. https://doi.org/10.1016/j.ecoinf.2016.08.003

Hartnett, E. & Rew, R. (2008). Experience with an enhanced netCDF data model and interface for scientific data access, 88th AMS Annual Meeting, 24th Conference on IIPS

Hermann, J. (2017). Meet Michaelangelo: Ubers Machine Learning Platform. Retrieved from https://eng.uber.com/michelangelo/

International DOI Foundation. (2011). DOI Handbook. http:// 10.1000/182 https://www.doi.org/hb.html

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the Reproducibility of Psychological Science. Journal of the American Statistical Association, 112(517), 1–10. https://doi.org/10.1080/01621459.2016.1240079

Kahn, R., & Wilensky, R. (2006). A framework for distributed digital object services. International Journal on Digital Libraries, 6(2), 115–123. https://doi.org/10.1007/s00799-005-0128-x

Kluyver, T., Ragan-kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., … Willing, C. (2016). Jupyter Notebooks—a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas, 87–90. https://doi.org/10.3233/978-1-61499-649-1-87

Martín-Sempere, M. J., Garzón-García, B., & Rey-Rocha, J. (2008). Team consolidation, social integration and scientists' research performance: An empirical study in the Biology and Biomedicine field. Scientometrics, 76(3), 457–482. https://doi.org/10.1007/s11192-007-1866-x

Mayernik, M. S., Gross, M. B., Corson-Rikert, J., Daniels, M. D., Johns, E. M., Khan, H., … Stott, D. (2016). Building Geoscience Semantic Web Applications Using Established Ontologies. Data Science Journal, 15(January 2016). https://doi.org/10.5334/dsj-2016-011

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., … Yarkoni, T. (2016). How open science helps researchers succeed. ELife, 5(JULY), 1–19. https://doi.org/10.7554/eLife.16800

Mesirov, J. P. (2010). Accessible, Reproducible Research. Science. Vol 327 (January), 415–416.

Michener, W. K. (2015). Ecological data sharing. Ecological Informatics, 29(P1), 33–44. https://doi.org/10.1016/j.ecoinf.2015.06.010

Morse, W. (2013). Integration of frameworks, theories and models across disciplines for effective cross-disciplinary communication. In M. O'Rourke, S. Crowley, S. D. Eigenbrode, and J. D. Wulfhorst, eds. Enhancing Communication and Collaboration in Interdisciplinary Research. Thousand Oaks, Calif.: Sage Publications.

Nicolescu, B. (2008). Transdisciplinarity : Theory and practice (Advances in systems theory, complexity, and the human sciences). Cresskill, NJ: Hampton Press.

National Institute for Standards and Technology (2007). Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. http://physics.nist.gov/Pubs/guidelines/contents. html

National Research Council. (2001). Under the Weather: Climate, Ecosystems, and Infectious Disease. The National Academies Press. https://doi.org/https://doi.org/10.17226/10025

Overeem, I., Berlin, M.M., & Syvitski, J.P.M. (2013). Strategies for integrated modeling: The community surface dynamics modeling system example. Environmental Modelling & Software, 39, 314-321.

Papazoglou, M. P., & Heuvel, W.-J. Van Den. (2006). Service-oriented design and development methodology. International Journal of Web Engineering and Technology, 2(4), 412. https://doi.org/10.1504/IJWET.2006.010423

Peckham, S.D., Kelbert, A., Hill, M.C., & Hutton, E.W.H. (2016). Towards uncertainty quantification and parameter estimation for Earth system models in a component-based modeling framework. Computers & Geosciences, pp152-161.

Pennington, D. (2016). A conceptual model for knowledge integration in interdisciplinary teams: orchestrating individual learning and group processes. Journal of Environmental Studies and Sciences, 6(2), 300–312. https://doi.org/10.1007/s13412-015-0354-5

Perez, Fernando, Granger, B. (2007). IPython : A System for Interactive Computing, 21–29. https://doi.org/10.1109/MCSE.2007.53

Pessoa R.M., Silva E., Van Sinderen M. (2008). Enterprise interoperability with SOA: a survey of service composition approaches. 2008 12th Enterprise Distributed Object Computing Conference Workshops, pp 238–251

Popper, K. (1992) The Logic of Scientific Discovery. Psychology Press. Philosophy – 513 pages (reprint from 1959).

Ram, K. (2013). Git can facilitate greater reproducibility and increased transparency in science. Source Code for Biology and Medicine, 8(1), 7. https://doi.org/10.1186/1751-0473-8-7

Reichman, O. J., Jones, M. B., & Schildhauer, M. P. (2011). Challenges and opportunities of open data in ecology. Science (New York, N.Y.), 331(6018), 703–705. https://doi.org/10.1126/science.1197962

Russell, A. W., Wickson, F., & Carew, A. L. (2008). Transdisciplinarity: Context, contradictions and capacity. Futures, 40(5), 460–472. https://doi.org/10.1016/j.futures.2007.10.005

Seamon, E., Gessler, P.E., Abatzogou, J.T., Mote, P.W., Lee, S.S. (2019a). Agricultural insurance loss analysis of the Pacific Northwest, USA: 2001 – 2015. Doctoral dissertation chapter 1, University of Idaho.

Seamon, E., Gessler, P.E., Abatzogou, J.T., Mote, P.W., Lee, S.S. (2019b). Regression based random forest modeling of inland pacific northwestern drought-related wheat insurance loss using time-lagged climate correlation matrix association. Doctoral dissertation chapter 2, University of Idaho.

Sedransk, N., Young, L. J., Kelner, K. L., Moffitt, R. a., Thakar, A., Raddick, J., … Spiegelman, C. (2010). Make Research Data Public? Not Always so Simple: A Dialogue for Statisticians and Science Editors. Statistical Science, 25(1), 41–50. https://doi.org/10.1214/10-STS320

Shahi, D. (2015). Apache Solr. A Practical Approach to Enterprise Search. ISBN 978-1-4842-1070-3

Shi, L., Jones, W. D., Jensen, R. V., Harris, S. C., Perkins, R. G., Goodsaid, F. M., … Tong, W. (2008). The balance of reproducibility, sensitivity, and specificity of lists of differentially expressed genes in microarray studies. BMC bioinformatics, 9 Suppl 9(Suppl 9), S10. doi:10.1186/1471-2105-9-S9-S10

Stark, P. B. (2018). No reproducibility without preproducibility. Nature, 557(7707), 613. https://doi.org/10.1038/d41586-018-05256-0

Stipelman, B.A., Hall, K.L., Zoss, A., Okamoto, J., Stokols, D., & Börrner, K. (2014). Mapping the impact of transdisciplinary research: A visual comparison of investigator-initiated and team-based tobacco use research publications. SciMed Central, Special Issue on Collaboration Science and Translational Medicine.

Stokols D., Hall K. & Vogel A. (2013). Transdisciplinary Public Health: Definitions, Core Characteristics, and Strategies for Success. In: Haire-Joshu D, McBride T, eds. Transdisciplinary Public Health: Research, Methods, and Practice. 1st ed. San Francisco, CA: Jossey-Bass; 2013:3-30.

Stodden, V., McNutt, M., Bailey, D.H., Deelman, E., Gil, Y., Hanson, B., . . .Taufer, M. (2016). Enhancing reproducibility for computational methods. Science, 354(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Stonebraker, M. (2009). Big Data Means at Least Three Different Things. MIT Computer Science and Artificial Intelligence Lab.

Taufer, M., Deelman, E., Ioannidis, J. P. A., Hanson, B., Stodden, V., Gil, Y., … Heroux, M. A. (2016). Enhancing reproducibility for computational methods. Science, 354(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Thornton, P. E., Running, S. W., & White, M. A. (1997). Generating surfaces of daily meteorological variables in complex terrain.pdf. Journal of Hydrology, 190, 214–251.

Treinish, L.A., & Gough, M.L. (1987). A software package for the data independent management of multi-dimensional data. EOS Trans Am Geophysical Union. Vol 68, No. 28.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. Science, 342(6157), 468–472. https://doi.org/10.1126/science.1240474

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Rennison, D. J. (2014). The availability of research data declines rapidly with article age. Current Biology : CB, 24(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Wandell, B. A., Rokem, A., Perry, L. M., Schaefer, G., & Dougherty, R. F. (2015). Data management to support reproducible research. Retrieved from http://arxiv.org/abs/1502.06900

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018. Retrieved from https://doi.org/10.1038/sdata.2016.18

Winn, J. (2013). Open data and the academy: an evaluation of CKAN for research data management. Iassist, 1–21. Retrieved from http://eprints.lincoln.ac.uk/9778/

Wuchty, S., Jones, B. & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. Plant Cell, 316(May), 2005–2008.

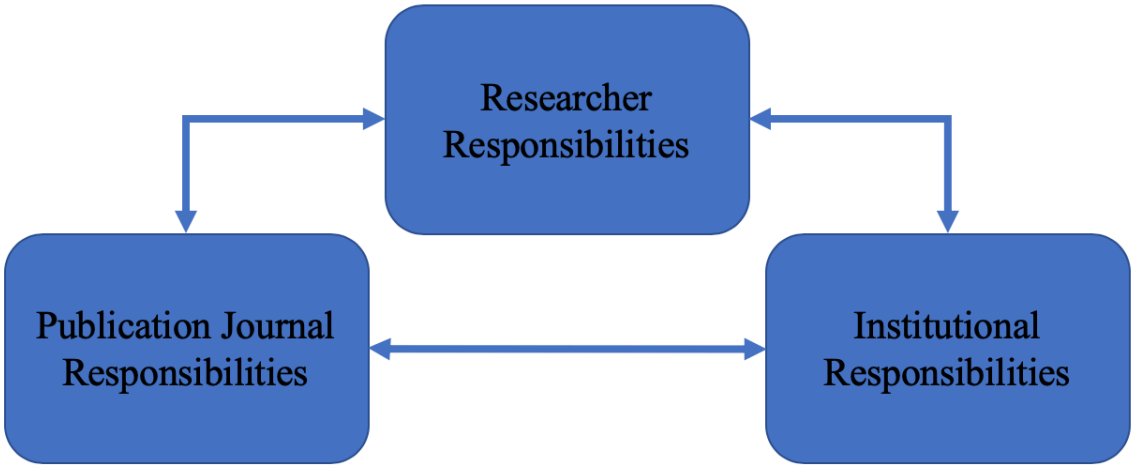
Xu, H., Li, S., Bai, Y., Dong, W., Huang, W., Xu, S., & Wu, T. (2019). Environmental Modelling & Software A collaborative analysis framework for distributed gridded environmental data. Environmental Modelling and Software, 111(June 2017), 324–339. https://doi.org/10.1016/j.envsoft.2018.09.007

Yale School Roundtable on Data and Code Sharing. (2010). Reproducible Research. Computing in Science & Engineering, 12(5), 8–13. https://doi.org/10.1109/mcse.2010.113

Zikopoulos P., DeRoos D., Parasuraman K., Deutsch T., Corrigan D., Giles J. (2013). Harness the Power of Big Data: The IBM Big Data Platform. ISBN:978-0-07180818-7.

Zimmerman, a. S. (2008). New Knowledge from Old Data: The Role of Standards in the Sharing and Reuse of Ecological Data. Science, Technology & Human Values, 33(5), 631–652. https://doi.org/10.1177/0162243907306704

Zender, C. S. (2006), netCDF Operators (NCO) for analysis of self-describing gridded geoscience data, Submitted to Environ. Modeling Software, available from http://dust.ess.uci.edu/ ppr/ppr Zen07.pdf. 1, 2.2, 2.4, 2.7, 3



Researcher Responsibilities	Institutional Responsibilities	Journal Responsibilities
Publish code and data in a location that is publicly accessible and persistent	Establish institutional hosting for data and analytics preservation	Require reproduction of results before publication
Uniquely identify versioning	Encourage and develop leadership to facilitate reproducibility standards	Require appropriate code and data citations through standardized citation mechanisms, such as Data Cite (http://thedata.org/citation/tech).
Use open licensing to facilitate code usage	Encourage researchers to use tools that embed code and data into publications	Require stable URLS for data and code for publication
When possible, use open access for publications	Fund data provenance and workflow sharing	
Publish code in non-proprietary formats		

Figure 3.1. Reproducibility responsibilities (Yale Law School Roundtable on Data and Code Sharing , 2010)

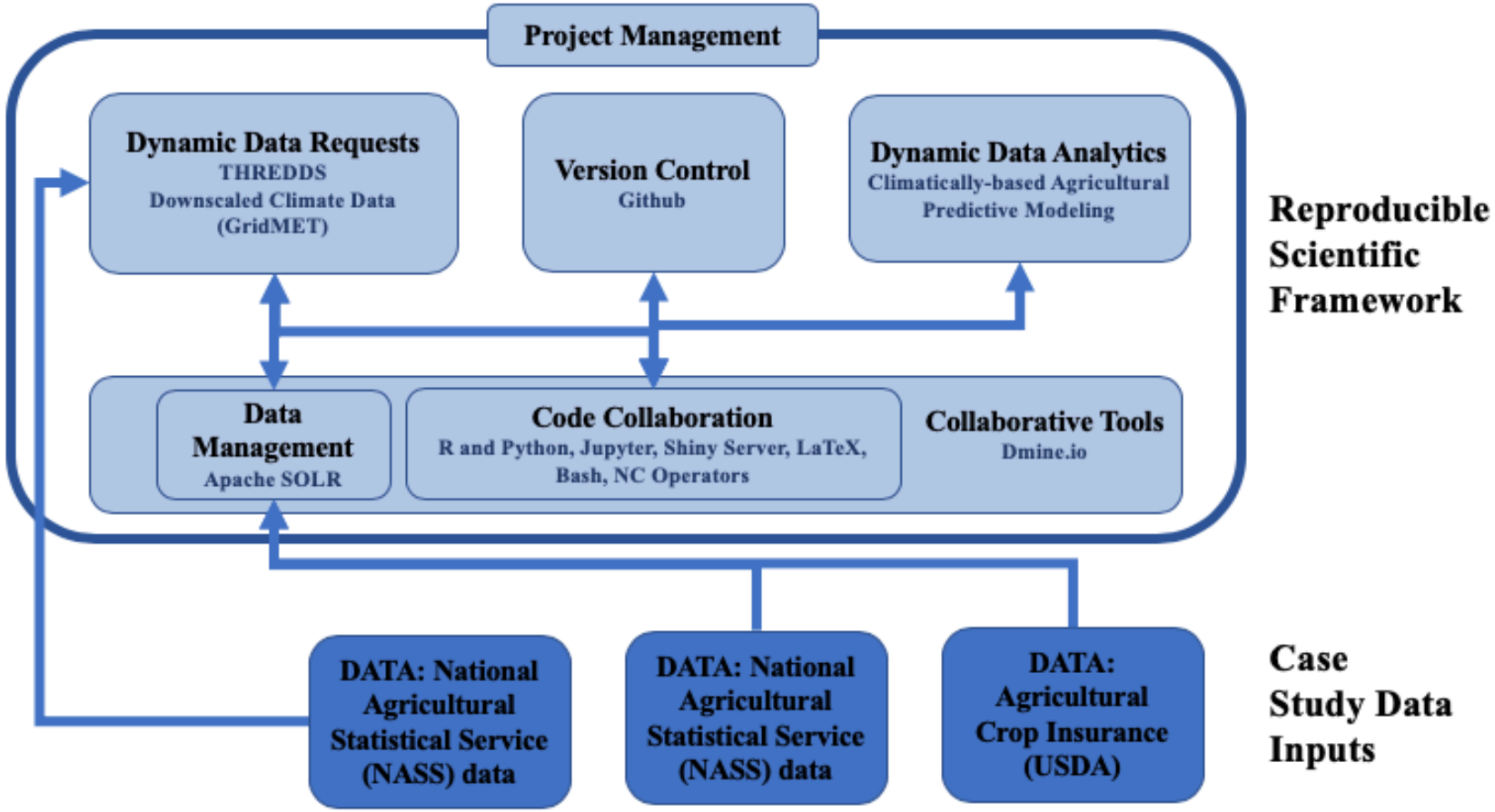


Figure 3.2. Reproducible Scientific Framework - Agriculture and Climate Case Study.

Gridded Climate Data Sources	Description
National Oceanic and Atmospheric Administration's (NOAA) National Center for Environmental Information (NCEI)	NOAA's NCEI hosts oceanic, atmospheric, and geophysical data. NCEI is the United States' leading agency for managing environmental information, and operates an extensive THREDDS repository of gridded climatic datasets: https://www.ncei.noaa.gov/thredds/model/model.html
National Aeronautical Space Agency (NASA) Distributed Active Archive Centers (DAACs)	NASA DAACs are part of the organizations' Earth Observing System Data and Information System (EOSDIS). DAACs provide access to a wide variety of earth-based data: https://earthdata.nasa.gov/eosdis/daacs https://search.earthdata.nasa.gov/search
United States Geological Survey (USGS) Data Portal	The USGS provides an extensive amount of data through their online portal and maintains a robust THREDDS repository for dynamic data access: https://www.usgs.gov/products/data-and-tools/data-and-tools-topics https://cida.usgs.gov/thredds/catalog.html
Australia's Commonwealth Scientific and Industrial Research Organisation (CSIRO) Data Portal	CSIRO is Australia's independent scientific research organization, and houses an array of climatic data sources: https://data.csiro.au/collections/
National Oceanic and Atmospheric Administration ERDDAP Servers	NOAA has developed an internal DAP compliant server called ERDDAP, which is extremely useful for extracting gridded array and table data. https://www.ncei.noaa.gov/erddap/index.html

Table 3.1. List of key climatic data repositories, many of which utilize software approaches to expose and enable access to data in gridded array formats.

Framework	Usage	Description	Case study application
1. Dynamic Data Requests	Thematic Realtime Environmental Distributed Data Services (THREDDS)	The THREDDS Data Server (TDS) is a web server that provides metadata and data access for scientific datasets, using OPeNDAP, OGC WMS and WCS, HTTP, and other remote data access protocols. THREDDS was developed and is supported by Unidata, a division of the University Corporation for Atmospheric Research (UCAR) and is sponsored by the National Science Foundation. (UCAR, 2019).	University of Idaho THREDDS catalog http://thredds.northwestknowledge.net
2. Version Control	Github	Version control provides an effective collaboration mechanism, particularly for code.	http://github.com/erichseamon/seamon_dissertation
3. Dynamic Data Analytics	RStudio's Shiny Server	Shiny Server is an R-based analytics server that allows for dynamic modeling generation.	Dmine Shiny Server Dashboards: http://dmine.io/dashboards
4. Data Management and Metadata	Apache Solr	Hyrax is a data server, developed to facilitate data access protocol (DAP) requests. Hyrax can be used in conjunction with THREDDS.	Dmine.io Solr instance: http://dmine.io/solr
5. Code Collaboration	Jupyter Notebooks, Rstudio Server, Rmarkdown	The Jupyter Project (Perez, 2007) provides collaborative research notebooks using Python, R, and other scientific languages. Rstudio Server permit R development from a web server framework.	Code and RMarkdown notebooks which reproduce analyses: http://github.com/erichseamon/seamon_dissertation
6. Collaborative Tools	Dmine.io web site	The Dmine.io web site provides a structure to present search, data, and analytics, as well as informational and instructional content.	Dmine.io web site: http://dmine.io
7. Project Management	Agile Project Management	Agile project management is centered round the idea of iterative development, where requirements and solutions evolve through collaboration between self-organizing cross-functional teams. methodologies.	Not implemented in case study

Table 3.2. Listing of components of the Reproducible Scientific Framework (RSF), along with the specific applications for our case study focusing on agriculture and climate relationships in the iPNW.

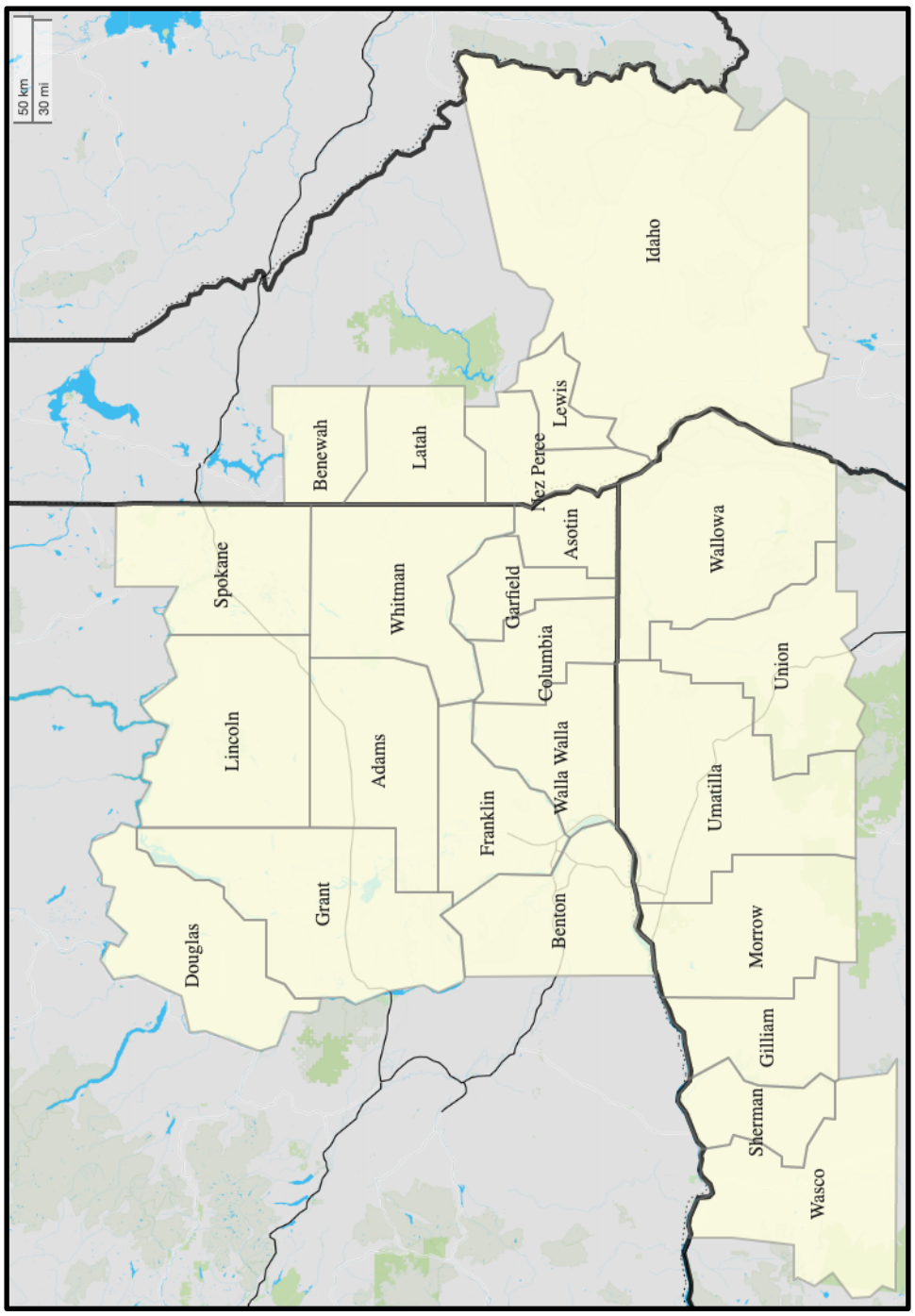


Figure 3.3. 24-county inland Pacific Northwest (iNPW) case study example area, which includes counties from Washington, Idaho, and Oregon.

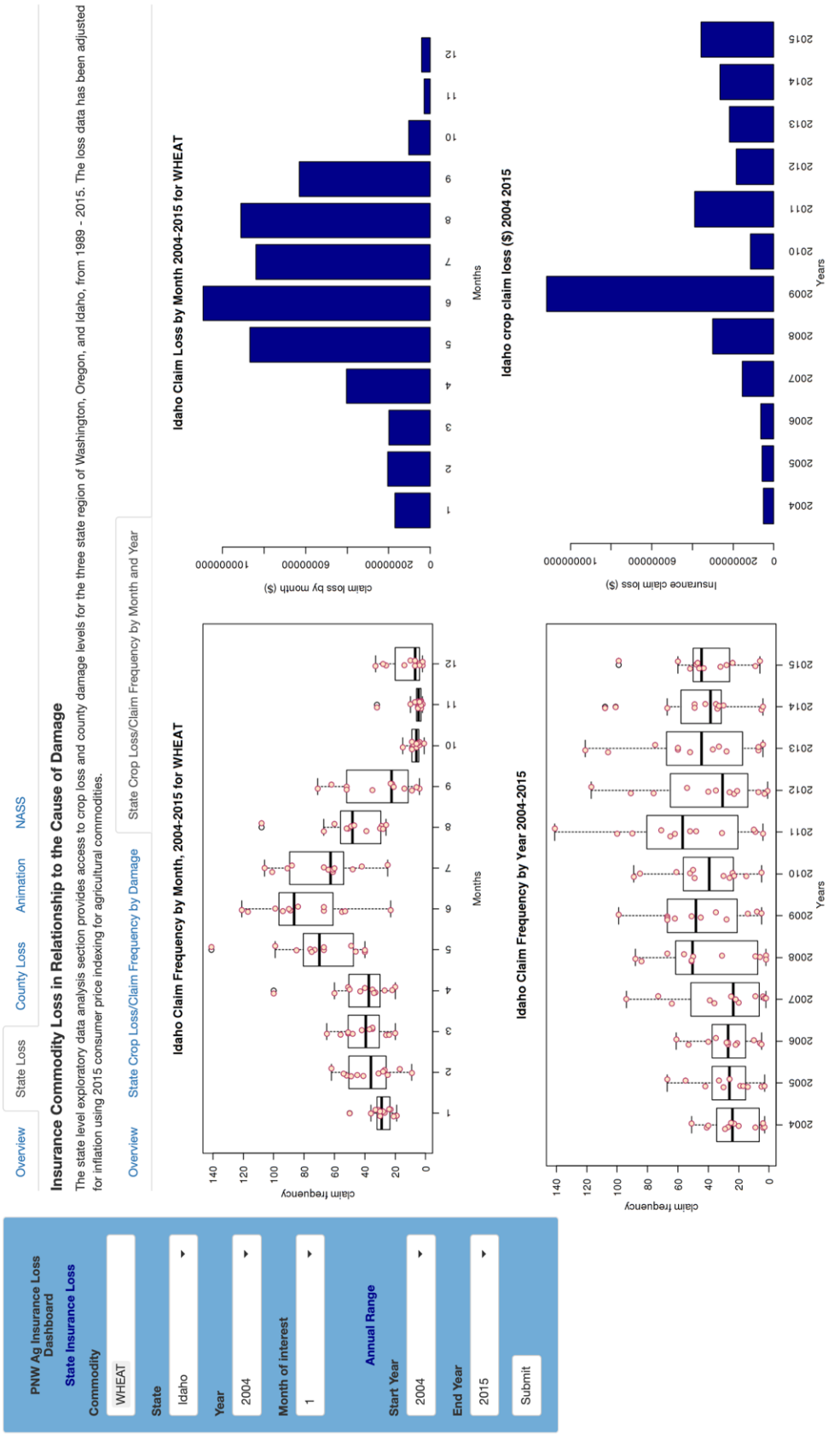


Figure 3.4. Example of exploratory data analysis (EDA) dashboard for examining agricultural insurance loss.

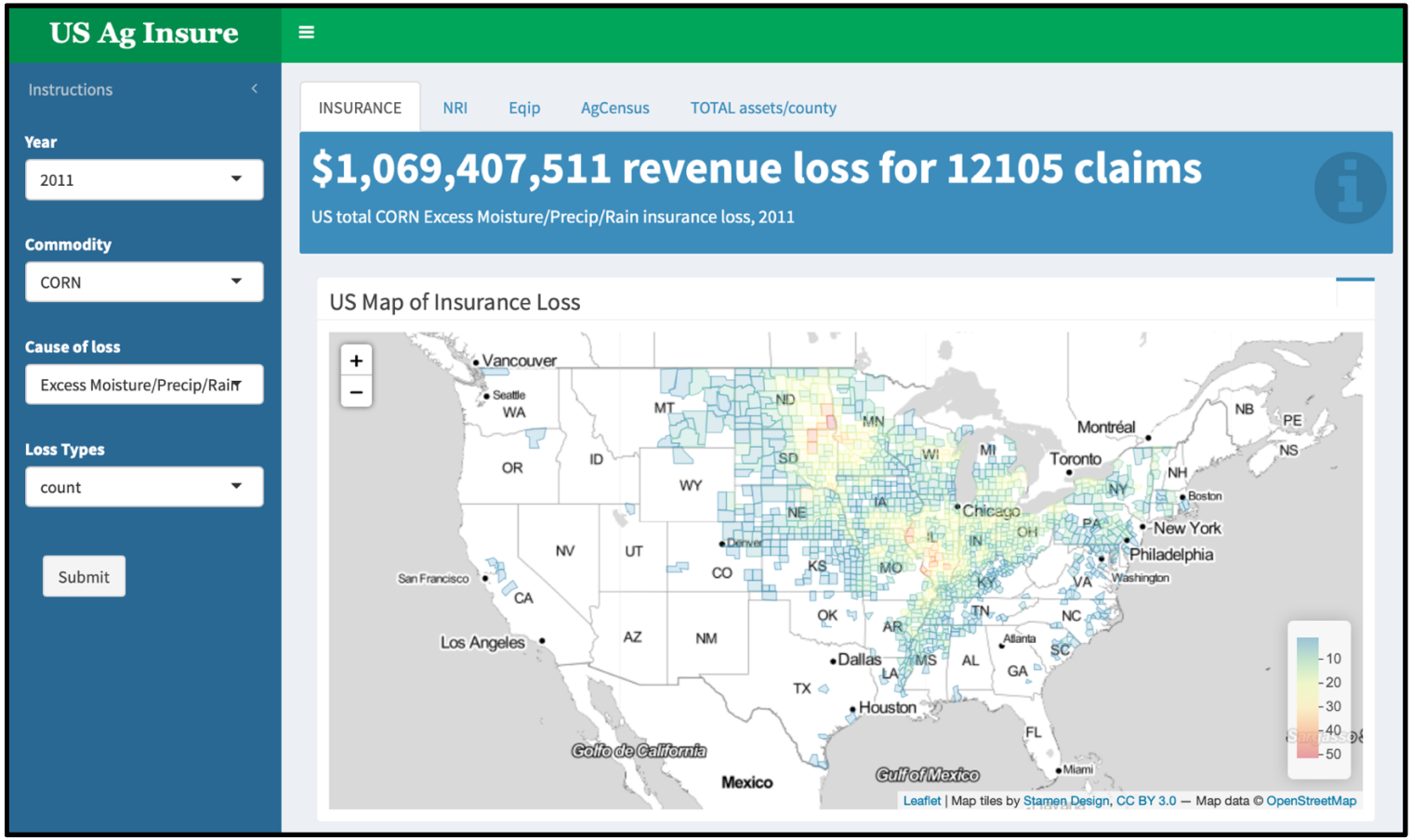


Figure 3.5. Example exploratory data analysis dashboard for nationwide agricultural insurance loss (http://dmine.io/dashboards).

Agricultural Prediction Dashboard: Insurance Loss Gradient Boosted Regression – iPNW Wheat Claims Due to Drought vs. Climate

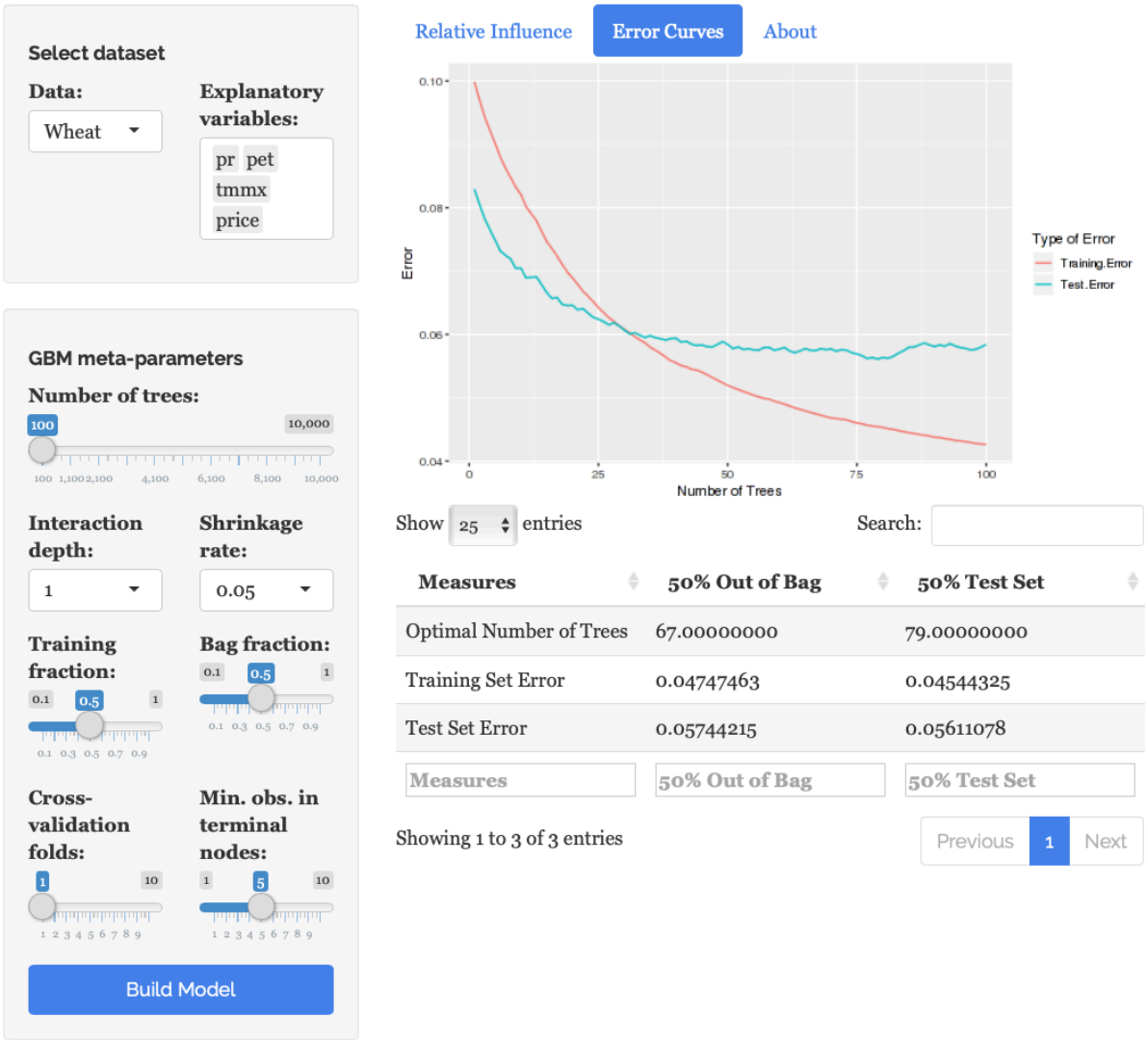


Figure 3.6. Example Shiny server predictive dashboard. This example dashboard, which is focused on the inland Pacific Northwest (iPNW), provides a predictive model to estimate agricultural insurance loss as compared to climatic variables and commodity pricing. The dashboard specifically uses a gradient boosted regression algorithm to estimate relative feature importance and error comparisons of test and train datasets (http://dmine.io/dashboards).

APPENDIX A: INSURANCE LOSS EXPLORATORY DATA ANALYSIS

Appendix A is an online RMarkdown generated HTML file which is associated with Chapter 1, and provides code and extended analytics output, describing crop insurance exploratory data analysis for the Pacific Northwest, as well as the inland Pacific Northwest.

Title	Online location
Appendix A: crop insurance exploratory data analysis	http://github.com/erichseamon/seamon_dissertation/ **Rmarkdown:**RMarkdown: seamon_dissertation_appendix_A.Rmd Html: seamon_dissertation_appendix_A.html File can be dynamically viewed at: http://erich.io/dissertation

APPENDIX B: INSURANCE LOSS PRINCIPAL COMPONENTS ANALYSIS

Appendix B is an online RMarkdown generated HTML file that is associated with Chapter 1 and provides code and extended analytics output with a focus on principal components analysis (PCA), for the differing factorial relationships for PNW and iPNW insurance loss (damage cause, year, county, and commodity type).

Title	Online location
Appendix B: principal components analysis (PCA)	http://github.com/erichseamon/seamon_dissertation/ **RMarkdown:** seamon_dissertation_appendix_B.Rmd Html: seamon_dissertation_appendix_B.html File can be dynamically viewed at: http://erich.io/dissertation

APPENDIX C: AGRICULTRUAL INSURANCE LOSS AND RELATIONSHIPS TO CLIMATE IN THE INLAND PACIFIC NORTHWEST

Appendix C is an online RMarkdown generated HTML file that is associated with Chapter 2 and provides code and extended analytics output related to wheat/drought insurance loss and the varying spatiotemporal relationships to climate. Appendix C documents the time-lagged climate correlation process and provides associated inline code.

Title	Online location
Appendix C: Insurance loss and relationships to climate	http://github.com/erichseamon/seamon_dissertation/ **RMarkdown:** seamon_dissertation_appendix_C.Rmd html: seamon_dissertation_appendix_C.html File can be dynamically viewed at: http://erich.io/dissertation

APPENDIX D: DATA AND CODE SOURCES

Appendix D contains online links to dissertation data sources related to all analyses.

Data Source	Online location
DATA: Data Loading	seamon_dissertation_dataload.R This script is used to load all datasets indicated below. The script can be run and will download all data to a /tmp directory on your local machine.
DATA: USDA/RMA Agricultural Insurance Loss	http://github.com/erichseamon/seamon_dissertation/data/RMA_originaldata/ Annual txt files containing individual insurance claims data from 1988 to 2015 (e.g. 1988.txt), for the entire United States. http://github.com/erichseamon/seamon_dissertation/data/RMA_csv/ Aggregated insurance claim files for the Pacific Northwest (Idaho, Oregon, and Washington), in comma separated file (csv) format. http://github.com/erichseamon/seamon_dissertation/data/RMA_Rda Insurance claim files in .Rda format, for the Pacific Northwest as well as for the entire United States
DATA: Wheat pricing	http://github.com/erichseamon/seamon_dissertation/data/wheat_prices

Data Source	Online location
DATA: Climatology	Gridded daily climate data for 14 variables (burn index, precipitation, wind speed, palmer drought severity index, minimum relative humidity, maximum relative humidity, minimum temperature, maximum temperature, solar radiation, wind direction, specific humidity, 100 hour fuel moisture, 1000 hour fuel moisture), aggregated at a monthly/county level, for Idaho, Washington, and Oregon – 1989 to 2015. http://github.com/erichseamon/seamon_dissertation/data/climatology http://github.com/erichseamon/seamon_dissertation/data/climate_matrices http://github.com/erichseamon/seamon_dissertation/data/climate_outputs http://github.com/erichseamon/seamon_dissertation/data/climate_correlations http://github.com/erichseamon/seamon_dissertation/data/climate_correlation_summaries
DATA: States	http://github.com/erichseamon/seamon_dissertation/data/states
DATA: Counties	http://github.com/erichseamon/seamon_dissertation/data/counties

APPENDIX E: CODE SOURCES

Appendix E contains online links to dissertation code sources related to all chapter analyses.

Code Source	Online location
CODE: Exploratory data analysis code and transformations (Chapter 1)	http://github.com/erichseamon/seamon_dissertation/appendices/appendix_A_code
CODE: Principal components analysis code (Chapter 1)	http://github.com/erichseamon/seamon_dissertation/appendices/appendix_B_code
CODE: Time-lagged climate correlation generation and regression random forest modeling (Chapter 2)	http://github.com/erichseamon/seamon_dissertation/appendices/appendix_C_code

APPENDIX F: EXPLORATORY DATA ANALYSIS DASHBOARD SOURCES

Appendix F contains online links to online exploratory data analysis dashboard sources, as discussed in Chapter 3.

Code Source	Description and online location
DASHBOARD: PNW Exploratory data analysis: agricultural insurance loss normal analysis (Chapter 3)	Shiny Server dashboard which compares spatial and temporal anomalies of insurance loss (by county and by month) https://dmine.io/ag-commodity-dashboard-ag-normals/
DASHBOARD : PNW Insurance loss county level comparison to climate (Chapter 3)	Shiny server dashboard which compares annual insurance loss by damage cause to annual climate variables. https://dmine.io/ag-commodity-loss-dashboard-climate-comparisons/
DASHBOARD : Nationwide agricultural insurance loss comparison with climate, by county (Chapter 3)	Shiny Server dashboard which compares United States insurance loss by county with differing climate variable totals. https://dmine.io/ag-commodity-loss-dashboard-nationwide-climate-data-by-county/
DASHBOARD : Nationwide agricultural insurance loss analysis (Chapter 3)	Shiny Server dashboard which analyzes agricultural insurance loss by county, year, commodity, and damage cause. https://dmine.io/agricultural-data-discovery-dashboard-nationwide/
DASHBOARD: PNW agricultural insurance loss analysis (Chapter 3)	Shiny Server dashboard which analyzes PNW insurance loss by state and county, with included animation. https://dmine.io/pnw-ag-insurance-loss-dashboard/

APPENDIX G: PREDICTIVE DASHBOARD SOURCES

Appendix G contains online links to online predictive dashboard sources, as discussed in Chapter 3.

Code Source	Description and online location
DASHBOARD: PNW Gradient Boosted Regression (Chapter 3)	Shiny Server predictive dashboard that uses a boosted regression approach to examine climate vs. wheat insurance loss. https://dmine.io/ag-gb-dashboard/
DASHBOARD : PNW regression, decision tree, and neural networking analysis (Chapter 3)	Shiny server dashboard which runs several analysis techniques to compare wheat insurance loss to climate variables (regression, decision tree analysis, and neural networking). https://dmine.io/ag-insurance-neural-network-dashboard/