Bayesian Inference of Severe Hail in Australia

Isabelle Greco, Steven Sherwood, Timothy Raupach, Gab Abramowitz
Climate Change Research Centre, University of New South Wales
ARC Centre of Excellence for Climate Extremes

Abstract

Severe hailstorms are responsible for some of the most costly insured weather events in Australia and can cause significant damage to homes, businesses, and agriculture. However their response to climate change remains uncertain, in large part due to the challenges of observing severe hailstorms. We propose a novel Bayesian approach which explicitly models known biases and uncertainties of current hail observations to produce more realistic estimates of severe hail risk from existing observations. Training this model on data from south-east Queensland, Australia, suggests that previous analyses of severe hail that did not account for this uncertainty may produce poorly calibrated risk estimates. Preliminary evaluation on withheld data confirms that our model produces well-calibrated probabilities and is applicable out of sample. Whilst developed for hail, we highlight also the generality of our model and its potential applications to other severe weather phenomena and areas of climate change adaptation and mitigation.

1 Introduction

On a continent regularly ravaged by bushfires, floods, and cyclones, storms producing hail greater than 2cm in diameter (hereafter severe hailstorms) over densely populated cities still stand among Australia's most costly insured weather events (McAneney et al., 2019; Insurance Council of Australia, 2022) and can seriously damage homes, businesses, and agriculture (e.g. Changnon (1992); Trapp et al. (2006)). To better prepare for these changing social and economic impacts as our climate warms, it is first necessary for us to know when and where severe hail risk is highest today (Púčik et al., 2019). However, hailstorms are only beginning to be resolved even in high-resolution climate models due to their small spatial and temporal scale and our incomplete knowledge of the relevant microphysics (Raupach et al., 2021; Mahoney et al., 2012; Brooks, 2013; Leslie et al., 2008). Hence, we must rely upon observations, which are themselves rare and biased, to understand severe hail risk (Brooks et al., 2003; Knight and Knight, 2001; Raupach et al., 2021; Schuster et al., 2005).

In situ severe hail observations (hereafter hail reports) are typically made by storm chasers and the general public. Thus, many hail events are not reported and hail reports are concentrated around densely-populated areas introducing a significant, temporally-varying, spatial bias into the dataset (Brimelow and Taylor, 2017; Allen and Karoly, 2014; Allen et al., 2011; Tuovinen et al., 2015). Hence, many analyses of hail risk employ radar data, which provide consistent spatial and temporal coverage (Warren et al., 2020; Brook et al., 2021). However, radar data alone is insufficient to definitively separate severe hail from hail less than 2cm in diameter and heavy rain (Knight and Knight, 2001; Brook et al., 2021). Whilst thresholds on various radar products exist which attempt to discriminate these two categories (e.g. Warren et al. (2020)), these thresholds a frequently selected from analysis of biased reports and do not account for uncertainty in the link between radar data and severe hail at the surface.

We take a different approach by developing a probabilistic Bayesian framework relating radar data to the probability of hail at the surface whilst accounting for chronic under-reporting of hail. Our approach facilitates the rigorous inclusion of prior knowledge and a

probabilistic characterisation of observational uncertainties. We apply the model to south-east Queensland, Australia, a known hail hotspot (Soderholm et al., 2019), although stress the model's generalisability.

2 Data and Bayesian model

We compile, for our model, a 0.25°, 6-hourly gridded dataset combining radar observations, hail reports, and population density estimates south-east Queensland (see Appendix A). From radar, we utilise the Maximum Expected Size of Hail (MESH) (Witt et al., 1998), a transformed, temperature-weighted, vertical integral of reflectivity, and take the maximum in each grid cell. Only data during Australia's hail season, taken conservatively to be September - April inclusive (Allen and Karoly, 2014; Allen et al., 2011; Schuster et al., 2005; Rasuly et al., 2015), are considered. Our models are trained on data from January 2010 - April 2015 inclusive and tested upon September 2015 - April 2016.

We derive our model in Appendix B. Crucially, for grid cell i at time t with associated observations $m_i^{(t)}$ and d_i of MESH and population density respectively, we model the probability of a report $R_i^{(t)}$ via the latent binary hail variable $H_i^{(t)}$ using logistic functions. Specifically, the likelihood of hail in a given cell is

$$\begin{aligned} P(R_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta) &= P(R_i^{(t)} = 1 \mid H_i^{(t)}, d_i, \Theta) P(H_i^{(t)} = 1 \mid m_i^{(t)}, \Theta) \\ &= \frac{1}{1 + \exp\left[-\left(\delta_1 f\left(d_i\right) + \delta_2\right)\right]} \cdot \frac{1}{1 + \exp\left[-\left(\mu_1 g\left(m_i^{(t)}\right) + \mu_2\right)\right]}, \end{aligned} \tag{1}$$

where f and g are transformations reducing the skew of the predictors (specified in Appendix B), and $\Theta = (\delta_1, \delta_2, \mu_1, \mu_2)$. This formulation assumes no false reports are made and that MESH has no bearing on the probability of hail being reported. We further assume conditional independence between spatiotemporal grid cells given the MESH and density observations. Justifications and proposals to relax these assumptions are detailed in Appendix B. Unlike in maximum likelihood estimation, we have the opportunity to specify priors upon Θ (see Appendix B) which incorporate scientific knowledge. Specifically, we enforce $\mu_1 > 0$ to reflect that higher MESH values represent stronger evidence of severe hail at the surface (Witt et al., 1998; Warren et al., 2020) and $\delta_1 > 0$ as reports are more likely in densely populated areas (Allen and Allen, 2016; Barras et al., 2019; Allen and Tippett, 2015). Bayesian inference also facilitates robust uncertainty quantification.

Our models are fitted using `RStan` (Stan Development Team, 2022a), an R interface to `Stan` (Stan Development Team, 2022b), itself an open-source implementation of the No-U-Turn Sampler (NUTS), a Hamiltonian Markov Chain Monte Carlo (MCMC) sampling algorithm (Betancourt, 2017). We utilised four chains in parallel to estimate Θ, each with 1000 burn-in samples which were discarded and 1000 retained samples. Analysis of the data and models was undertaken largely in R (R Core Team, 2022) employing heavily `bayesplot` (Gabry and Mahr, 2022) and the `tidyverse` (Wickham et al., 2019). NUTS diagnostics, the $\hat{R}$ statistic, and the effective sample size N_{eff} were used to evaluate numerical performance (Gelman et al., 2004; Betancourt, 2017). We assessed model fit via posterior predictive checking (Gelman et al., 2004), reliability diagrams (Bröcker and Smith, 2007), and the ignorance score (Roulston and Smith, 2002).

3 Results

Fitting this model resulted in $\hat{R} < 1.002$ and $N_{\text{eff}} > 1500$ for Θ indicating excellent Markov chain mixing and sufficient quasi-independent draws for Bayesian inference (Gelman et al., 2004). Figure 1 reveals the fitted functions in Equation 1 as calculated from the posterior samples. Particularly interesting is the low probabilities of hail at moderate MESH values. Typical radar-based analyses of Australian hail take MESH values above a threshold of 21mm to 32mm to indicate severe hail (Brook et al., 2021). However, at the 95% confidence level, our model suggests the probability of hail is between $[0.0449, 0.1620]$ at 21mm MESH

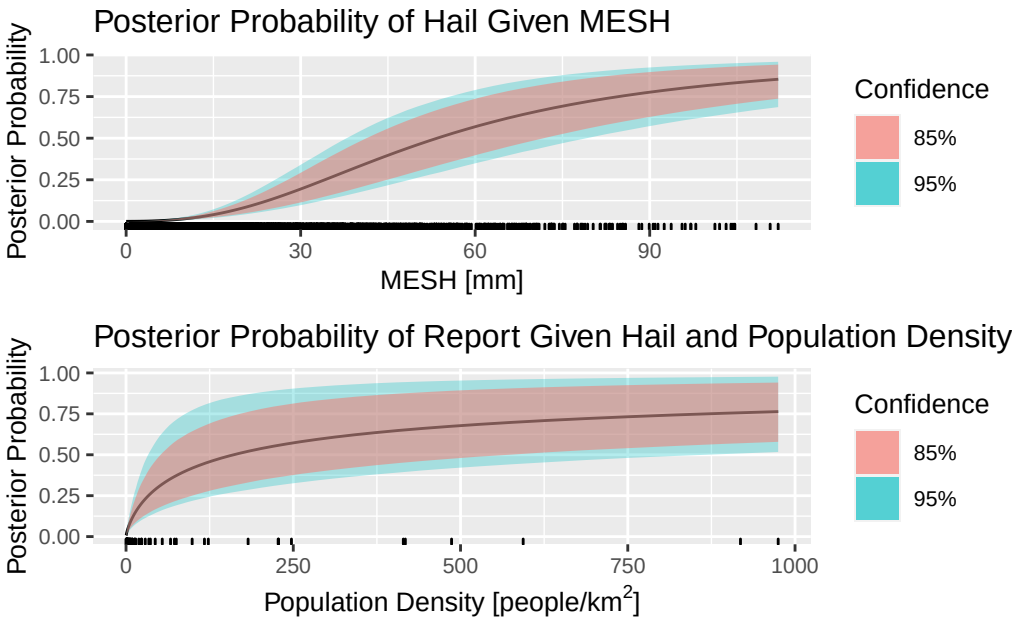

Figure 1: The posterior probability of (top) hail given MESH and (bottom) the conditional probability of a report given hail fell as a function of the population density. Both curves are logistic functions in the transformed predictors $f(\text{MESH})$ and $g(\text{density})$.

and $[0.1118, 0.3912]$ at 32mm. Hence, previous analyses of Australian hail may be poorly calibrated and may inaccurately estimate current and future hail risk.

Although our model yields full distributions, and a reliability diagram suggests the probabilities are well-calibrated (see Appendix D), we take the expected number of hail days as a point estimate in Figure 2 when evaluating the model's performance on the test set, September 2015 - April 2016. Note that when MESH alone is utilised, the number of hail days during the season is highly sensitive to the threshold chosen. Equation 1 avoids this issue by modelling the probability of severe hail given a MESH observation as a continuous function, rather than treating MESH as a binary indicator. Moreover, comparison with the analysis derived solely from reports, which are concentrated around the state capital Brisbane and major highways, reveals that our model was able to use the information contained within the reports without replicating the same biases.

Quantitatively, the mean ignorance, an information theoretic metric to assess probabilistic forecasts[1] (Roulston and Smith, 2002), for the expected reporting probability is 0.0218 on the training set and 0.0217 on the test set, suggesting our model generalises well across seasons. As a reference, we consider the performance of a naive probabilistic forecast which constantly predicts a reporting probability equal to the proportion of MESH observations with reports in the training set. This naive forecast gives a mean ignorance of 0.0415 on the training set and 0.0388 on the test set, indicating the value added by our model.

4 Sensitivity Testing

Similar models applied to tornado reports (e.g. Potvin et al. (2019; 2022)) have encountered solution non-uniqueness and parameter confounding. Specifically, models struggled to differentiate high reporting rates and low occurrence rates from low reporting rates and high occurrence rates, as both can produce the same number of reports. We simulated data from each of these two cases using distributions which produced the same expected number of reports and fitted our model to each of the 100 simulated data sets using strongly in-

[1]Note that in our case using the ignorance score and the likelihood give the same results under the assumption of temporally uncorrelated errors.

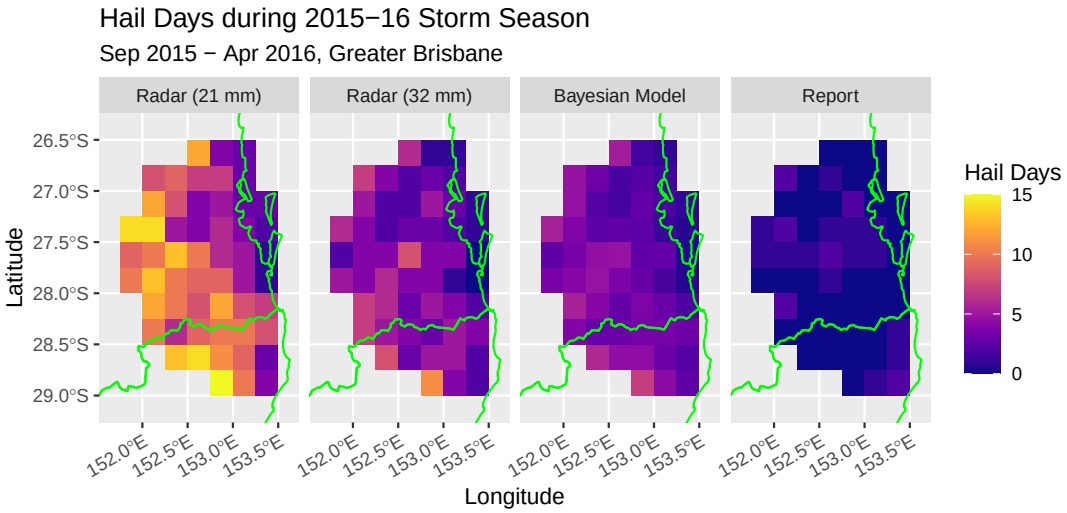

Figure 2: The expected number of hail days from the Bayesian model applied to the September 2015 - April 2016 in comparison with two example radar thresholds and reports.

formative, weakly informative, and misleading priors (see Appendix C for further details). We found that misleading priors may induce small bias, but in every case our model was able to consistently and accurately differentiate the two regimes and avoided solution non-uniqueness. We attribute our success to the utilisation of MESH as a predictor rather than employing a spatially random process of tornadogenesis (Potvin et al., 2019; 2022).

We also verified model performance when the functions generating the data were not logistic functions, contrary to our model's assumptions (Equation 1). Whilst logistic functions were chosen due to their flexibility, these less idealised tests assisted in determining the model's robustness to real observations. We also tested alternative distributions of the predictors to assess the applicability of our model to unobserved data. Our experimentation with other functions and predictor distributions (not included in this paper), suggested that the model is most accurate when the predictors are uniformly distributed on a bounded range and the probability to be estimated can be feasibly approximated by a logistic function. This result motived our choices of f and g in Equation 1 as we selected functions which made the distribution of the transformed variables more uniform (see Appendix B).

5 Conclusions

In order to better understand observations of severe hail in Australia, and ultimately better characterise the current and future risk of these hazards, we have developed a novel Bayesian model for severe hail at the surface. Our model accounts for uncertainty and biases in two key data sources, radar and hail reports, and outputs probabilistic estimates of hail risk with associated uncertainties. Hence, our model enables decision-makers to more accurately assess hail risk and whilst being more aware of the uncertainty in our estimates. Our model's results in south-east Queensland suggest that conventional estimates using radar or reports alone may be biased. Furthermore, our preliminary model evaluation suggests our model has can generalise across hail seasons. Finally, our model is generalisable: the mathematics would not change if we were to consider other severe hail proxies at different times and places nor other severe weather reports like extreme wind or rain. Applications are also possible beyond meteorology to other situations where reports of an event are only received in the event's presence rather than its absence and the absence of a report does not guarantee non-occurrence of the event.

References

John T. Allen and Edwina R. Allen. A review of severe thunderstorms in Australia. *Atmospheric Research*, 178-179:347–366, 2016. ISSN 0169-8095. doi: 10.1016/j.atmosres.2016.03.011.

John T. Allen and David J. Karoly. A climatology of Australian severe thunderstorm environments 1979–2011: inter-annual variability and ENSO influence. *International Journal of Climatology*, 34(1):81–97, 2014. doi: 10.1002/joc.3667.

John T. Allen and Michael K. Tippett. The characteristics of United States hail reports: 1955-2014. *E-Journal of Severe Storms Meteorology*, 10(3):1–15, 2015.

John T. Allen, David J. Karoly, and Graham A. Mills. A severe thunderstorm climatology for Australia and associated thunderstorm environments. *Australian meteorological and oceanographic journal*, 61(3):143–158, 2011. ISSN 1836-716X.

Australian Bureau of Statistics. Population estimates by SA2 (ASGS2021), 2001 to 2021 [data set], 2021. URL `abs.gov.au/statistics/people/population/regional-population/2021`.

Australian Bureau of Statistics. Australian population grid 2021 in ESRI grid format [dataset], 2022. URL `abs.gov.au/statistics/people/population/regional-population/`.

Hélène Barras, Alessandro Hering, Andrey Martynov, Pascal-Andreas Noti, Urs Germann, and Olivia Martius. Experiences with >50,000 crowdsourced hail reports in Switzerland. *Bulletin of the American Meteorological Society*, 100(8):1429–1440, 08 2019. doi: 10.1175/BAMS-D-18-0090.1.

Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*, 2017.

Scott F. Blair, Derek R. Deroche, Joshua M. Boustead, Jared W. Leighton, Brian L. Barjenbruch, and William P Gargan. A radar-based assessment of the detectability of giant hail. *E-Journal of Severe Storms Meteorology*, 6(7):1 – 30, 2011.

Julian C. Brimelow and Neil M. Taylor. Verification of the MESH product over the Canadian prairies using a high-quality surface hail report dataset sourced from social media. In *38th Conference on Rediative Meteorology*, Organised Convection and Severe Phenomena, 2017.

Jochen Bröcker and Leonard A. Smith. Increasing the reliability of reliability diagrams. *Weather and Forecasting*, 22(3):651 – 661, 2007. doi: 10.1175/WAF993.1.

Jordan P. Brook, Alain Protat, Joshua Soderholm, Jacob T. Carlin, Hamish McGowan, and Robert A. Warren. Hailtrack—improving radar-based hailfall estimates by modeling hail trajectories. *Journal of Applied Meteorology and Climatology*, 60(3):237 – 254, 2021. doi: 10.1175/JAMC-D-20-0087.1.

Harold E. Brooks. Severe thunderstorms and climate change. *Atmospheric Research*, 123: 129–138, 2013. ISSN 0169-8095. doi: 10.1016/j.atmosres.2012.04.002.

Harold E. Brooks, James W. Lee, and Jeffrey P. Craven. The spatial distribution of severe thunderstorm and tornado environments from global reanalysis data. *Atmospheric Research*, 67-68:73–94, 2003. ISSN 0169-8095. doi: 10.1016/S0169-8095(03)00045-0.

Bureau of Meteorology. Radar site information. URL `bom.gov.au/australia/radar/about/radar_site_info.shtml`.

Stanley A. Changnon. Temporal and spatial relations between hail and lightning. *Journal of Applied Meteorology and Climatology*, 31(6):587 – 604, 1992. doi: 10.1175/1520-0450(1992)031<0587:TASRBH>2.0.CO;2.

John L. Cintineo, Travis M. Smith, Valliappa Lakshmanan, Harold E. Brooks, and Kiel L. Ortega. An objective high-resolution hail climatology of the contiguous united states. *Weather and Forecasting*, 27(5):1235 – 1248, 2012. doi: 10.1175/WAF-D-11-00151.1.

Andrew J. Dowdy, Joshua Soderholm, Jordan Brook, Andrew Brown, and Hamish McGowan. Quantifying hail and lightning risk factors using long-term observations around Australia. *Journal of Geophysical Research: Atmospheres*, 125(21):2020JD033101, 2020. doi: 10.1029/2020JD033101.

Jonah Gabry and Tristan Mahr. bayesplot: Plotting for Bayesian models, 2022.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Texts in Statistical Science. Chapman & Hall/CRC, 2nd edition, 2004.

Will H. Hand and Gennaro Cappelluti. A global hail climatology using the UK Met Office convection diagnosis procedure (CDP) and model analyses. *Meteorological Applications*, 18(4):446–458, 2011. doi: 10.1002/met.236.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: 10.1002/qj.3803.

Insurance Council of Australia. Insurance catastrophe resilience report 2021-22. Catastrophe report, Insurance Council of Australia, 2022.

Donald L. Kelly, Joseph T. Schaefer, and Charles A. Doswell. Climatology of nontornadic severe thunderstorm events in the united states. *Monthly Weather Review*, 113(11):1997–2014, 1985. doi: 10.1175/1520-0493(1985)113¡1997:CONSTE¿2.0.CO;2.

Charles A. Knight and Nancy C. Knight. Hailstorms. In Charles A. Doswell III, editor, *Severe Convective Storms*, volume 28 of *Meteorological Monographs*, chapter 6, pages 223–254. American Meteorological Society, 45 Beacon Street, Boston, Massachusetts, 2001.

Lance M. Leslie, Mark Leplastrier, and Bruce W. Buckley. Estimating future trends in severe hailstorms over the Sydney basin: A climate modelling study. *Atmospheric Research*, 87 (1):37–51, 2008. ISSN 0169-8095. doi: 10.1016/j.atmosres.2007.06.006.

Kelly Mahoney, Michael A. Alexander, Gregory Thompson, Joseph J. Barsugli, and James D. Scott. Changes in hail and flood risk in high-resolution simulations over Colorado's mountains. *Nature Climate Change*, 2(2):125–131, 2012. doi: 10.1038/nclimate1344.

John McAneney, Benjamin Sandercock, Ryan Crompton, Thomas Mortlock, Rade Musulin, Roger Pielke Jr, and Andrew Gissing. Normalised insurance losses from Australian natural disasters: 1966–2017. *Environmental Hazards*, 18(5):414–433, 2019. doi: 10.1080/17477891.2019.1609406.

Elisa M. Murillo and Cameron R. Homeyer. Severe hail fall and hailstorm detection using remote sensing observations. *Journal of Applied Meteorology and Climatology*, 58(5):947 – 970, 2019. doi: 10.1175/JAMC-D-18-0247.1.

Corey K. Potvin, Chris Broyles, Patrick S. Skinner, Harold E. Brooks, and Erik Rasmussen. A Bayesian hierarchical modeling framework for correcting reporting bias in the U.S. tornado database. *Weather and Forecasting*, 34(1):15 – 30, 2019. doi: 10.1175/WAF-D-18-0137.1.

Corey K. Potvin, Chris Broyles, Patrick S. Skinner, and Harold E. Brooks. Improving estimates of U.S. tornado frequency by accounting for unreported and underrated tornadoes. *Journal of Applied Meteorology and Climatology*, 61(7):909 – 930, 2022. doi: 10.1175/JAMC-D-21-0225.1.

Tomáš Púčik, Christopher Castellano, Pieter Groenemeijer, Thilo Kühne, Anja T. Rädler, Bogdan Antonescu, and Eberhard Faust. Large hail incidence and its economic and societal impacts across Europe. *Monthly Weather Review*, 147(11):3901 – 3916, 2019. doi: 10.1175/MWR-D-19-0204.1.

QGIS Development Team. *GQIS Geographical Information System*. QGIS Association, 2022. URL `qgis.org`.

R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

A. A. Rasuly, K. K. W. Cheung, and B. McBurney. Hail events across the greater metropolitan severe thunderstorm warning area. *Natural Hazards and Earth System Sciences*, 15 (5):973–984, 2015. doi: 10.5194/nhess-15-973-2015.

Timothy H. Raupach, Olivia Martius, John T. Allen, Michael Kunz, Sonia Lasher-Trapp, Susanna Mohr, Kristen L. Rasmussen, Robert J. Trapp, and Qinghong Zhang. The effects of climate change on hailstorms. *Nature Reviews Earth & Environment*, 2(3):213–226, 2021. doi: 10.1038/s43017-020-00133-9.

Mark S. Roulston and Leonard A. Smith. Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6):1653 – 1660, 2002. doi: 10.1175/1520-0493(2002)130<1653:EPFUIT>2.0.CO;2.

Elena Saltikoff, Jari-Petteri Tuovinen, Janne Kotro, Timo Kuitunen, and Harri Hohti. A climatological comparison of radar and ground observations of hail in Finland. *Journal of Applied Meteorology and Climatology*, 49(1):101 – 114, 2010. doi: 10.1175/2009JAMC2116.1.

Sandra S. Schuster, Russell J. Blong, and Milton S. Speer. A hail climatology of the greater Sydney area and New South Wales, Australia. *International Journal of Climatology*, 25 (12):1633–1650, 2005. doi: 10.1002/joc.1199.

Joshua S. Soderholm, Hamish McGowan, Harald Richter, Kevin Walsh, Tammy Weckwerth, and Matthew Coleman. The coastal convective interactions experiment (CCIE): Understanding the role of sea breezes for hailstorm hotspots in eastern Australia. *Bulletin of the American Meteorological Society*, 97(9):1687 – 1698, 2016. doi: 10.1175/BAMS-D-14-00212.1.

Joshua S. Soderholm, Kathryn I. Turner, Jordan P. Brook, Tony Wedd, and Jeffery Callaghan. High-impact thunderstorms of the Brisbane metropolitan area. *Journal of Southern Hemisphere Earth Systems Science*, 69(1):239–251, 2019.

Joshua S. Soderholm, V. Louf, J. Brook, A. Protat, and R. Warren. Australian operational weather radar level 2 dataset. *National Computing Infrastructure*, 2022. doi: 10.25914/JJWZ-0F13.

Stan Development Team. RStan: the R interface to Stan. R package version 2.21.5, 2022a.

Stan Development Team. Stan modeling language users guide and reference manual. mc-stan.org, 2022b.

Gregor Stržinar and Gregor Skok. Comparison and optimization of radar-based hail detection algorithms in slovenia. *Atmospheric Research*, 203:275–285, 2018. ISSN 0169-8095. doi: 10.1016/j.atmosres.2018.01.005.

Mateusz Taszarek, John T. Allen, Pieter Groenemeijer, Roger Edwards, Harold E. Brooks, Vanna Chmielewski, and Sven-Erik Enno. Severe convective storms across Europe and the United States. Part I: Climatology of lightning, large hail, severe wind, and tornadoes. *Journal of Climate*, 33(23):10239 – 10261, 2020. doi: 10.1175/JCLI-D-20-0345.1.

Robert J. Trapp, Dustan M. Wheatley, Nolan T. Atkins, Ronald W. Przybylinski, and Ray Wolf. Buyer beware: Some words of caution on the use of severe wind reports in postevent assessment and research. *Weather and Forecasting*, 21(3):408–415, 2006. doi: 10.1175/WAF925.1.

Jari-Petteri Tuovinen, Jenni Rauhala, and David M. Schultz. Significant-hail-producing storms in Finland: Convective-storm environment and mode. *Weather and Forecasting*, 30(4):1064 – 1076, 2015. doi: 10.1175/WAF-D-14-00159.1.

Robert A. Warren, Hamish A. Ramsay, Steven T. Siems, Michael J. Manton, Justin R. Peter, Alain Protat, and Anu Pillalamarri. Radar-based climatology of damaging hailstorms in Brisbane and Sydney, Australia. *Quarterly Journal of the Royal Meteorological Society*, 146(726):505–530, 2020. doi: 10.1002/qj.3693.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welbome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019.

Earle Williams, Bob Boldi, Anne Matlin, Mark Weber, Steve Hodanish, Dave Sharp, Steve Goodman, Ravi Raghavan, and Dennis Buechler. The behavior of total lightning activity in severe Florida thunderstorms. *Atmospheric Research*, 51(3):245–265, 1999. ISSN 0169-8095. doi: 10.1016/S0169-8095(99)00011-3.

Arthur Witt, Michael D. Eilts, Gregory J. Stumpf, J. T. Johnson, E. De Wayne Mitchell, and Kevin W. Thomas. An enhanced hail detection algorithm for the WSR-88D. *Weather and Forecasting*, 13(2):286 – 303, 1998. doi: 10.1175/1520-0434(1998)013<0286:AEHDAF>2.0.CO;2.

Appendices

A Gridded Data

We utilise radar observations from the Australian Bureau of Meteorology's (the Bureau's) Mount Staplyton S-band Doppler radar (ID: 66), situated on an isolated hill south-east of central Brisbane (Bureau of Meteorology). We take radar observations over land within approximately 120km of the radar itself, yielding a circular region of radar observations over land between 26.50°S and 29.00°S and 151.75°E and 153.75°E. This swath is slightly less than the range at which radar can retrieve data but we found radar retrievals beyond this point to be of lower quality, likely due to beam broadening and geographical obstructions (Bureau of Meteorology; Cintineo et al., 2012; Blair et al., 2011). It is also common to not employ data measured very close to the radar due to the cone of silence: Australian radars do not scan overhead (Soderholm et al., 2016; Blair et al., 2011). However, given we are employing this data at 6-hourly resolution on a 0.25° spatial grid (approximately 25km × 25km), any storm directly over the radar would very likely move away from radar and into a position within that grid cell area where it could be scanned. Hence we did not consider it necessary to exclude grid cells close to the radar from our analysis.

There are many possible indices derived from radar which can be used as proxies of varying skill for the presence of (severe) hail (Knight and Knight, 2001; Saltikoff et al., 2010; Blair et al., 2011; Brook et al., 2021). Whilst metrics derived from dual-polarised (DP) radars are typically more skilful, as they can distinguish oblate raindrops from quasi-spherical

hailstones (Murillo and Homeyer, 2019), DP radars are uncommon in Australia; the Mount Staplyton radar is itself single-polararised (Soderholm et al., 2022). Single-polarised indices are therefore our best choice if we wish our methods to be applicable widely in Australia. Arguably the most popular such hail index is the maximum expected size of hail (MESH) derived by Witt et al. (1998). Whilst generally unable to accurately determine the maximum expected size of hail (Witt et al., 1998; Warren et al., 2020; Cintineo et al., 2012; Brimelow and Taylor, 2017; Murillo and Homeyer, 2019), MESH, a transformation of a temperature weighted vertical integral of the effective reflectivity factor, is commonly used in operational forecasting as an indicator of severe hail [@warren20; @brimelow17]. MESH is provided within the cleaned and pre-processed radar retrievals supplied by the Bureau (Soderholm et al., 2022). The temperature required to weight the integral is derived from hourly ERA5 reanalysis data (Soderholm et al., 2022; Hersbach et al., 2020). We utilise the maximum MESH, where available, over the spatiotemporal grid cell to capture the storm's greatest severity as well as the drift of hailstones from the swath where they were measured (Brimelow and Taylor, 2017), but acknowledge the possibility of many other possible and more nuanced choices (like that of Warren et al. (2020) or Brook et al. (2021)).

The ABS provides freely available estimated resident population (ERP) data at a variety of resolutions between years 2001 to 2021 (Australian Bureau of Statistics, 2021). Although since updated, at the time these results were generated the only freely available gridded product gave the 2021 ERP at 1km^2 resolution (Australian Bureau of Statistics, 2022) which we transform from the Australian Albers coordinate reference system to a Mercartor grid using QGIS (QGIS Development Team, 2022). From here it was simple to, again in QGIS, take areal averages to coarsen this data to $0.25°$resolution. We note that the distribution of population in each grid cell is strongly skewed, and so, from hereon, utilise the logarithm of the population density. Given Brisbane's rapid population growth from 2010 to 2021 (Australian Bureau of Statistics, 2021), utilising this older data may introduce some error, and in future iterations of this work will update to use the newly available 2015 data. A further potential issue is the difficulty of defining the average population density of a coastal grid cell containing ocean. We opted to treat the ocean as unpopulated land and then averaged the density over this area, but do acknowledge that this approach may falsely deflate the population density of densely-populated coastal regions.

Hail reports were drawn from the Severe Storms Archive (SSA) which, as of 1987, is operated by the Bureau's Severe Weather Forecasting Section (Schuster et al., 2005; Hand and Cappelluti, 2011). Whilst the SSA is nominally still operational, the number of hail reports since the 2015-2016 storm season has decreased to near zero despite a spate of known severe events reported in the media and hence we do not attempt to utilise the SSA beyond this season. Report quality is also an issue: Schuster et al. (2005) found that a reliable time of event was only available for half the reports in the database. We manually investigated the timings of events reported before 1200 local time (Australian Eastern Standard Time, AEST) as there is a known peak in hail activity during the late afternoon and low activity in the morning (Dowdy et al., 2020; Taszarek et al., 2020; Schuster et al., 2005; Rasuly et al., 2015). We found that many of these morning reports were simply assigned the default time of 0000 UTC (1000 AEST). This default value was particularly common prior to 2010, prompting us to ignore these older reports, and manually examine and assign to the correct temporal bin the remaining 12 cases. Correction was informed by radar images, comments associated with the report, and contemporary news articles. We also noted from the comments some events before 1200 AEST indicated that these vents did take place in the afternoon, but the time was not correctly converted to UTC. We then made this conversion. These tow cases eliminated all reports in our study region before 1200 AEST. Whilst spatial errors are also possible (Schuster et al., 2005; Rasuly et al., 2015), and we corrected the most egregious errors (e.g. $0°$N, $0°$E), we did not investigate smaller errors further as we used coarse spatial bins that we effectively mitigated inaccuracies. In rare cases where two or more reports fall in the same spatiotemporal bin, the report corresponding to the largest hail size was retained.

B Model Derivation

We firstly assumed that if MESH was not recorded as there was simply no reflectivity in the region of probable hail growth (Witt et al., 1998), then the probability of hail falling was zero. Whilst imperfect, as MESH could be missing due to radar malfunction, beam blocking, or beam broadening (Stržinar and Skok, 2018) and so we may miss a small number of hail events, this assumption enables us to focus our analysis upon moments where hail was most plausible. Consider then grid cell i at time t, and associated available MESH observation $m_i^{(t)}$, population density d_i, and a binary indicator $R_i^{(t)}$ of a hail report being made. Note we considered only the existence of a hail report, rather than the information contained therein, due to the approximate and inconsistent nature of supplementary reporting information. Our variable of interest is the unobserved binary indicator of severe hail in cell i at time t, $H_i^{(t)}$. We cannot model $H_i^{(t)}$ directly and so we model it via the observed reports

$$P(R_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta)$$

where Θ contains the model parameters. We now introduce hail via $H_i^{(t)}$ obtaining

$$\begin{aligned} P(R_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta) = P(R_i^{(t)} = 1, H_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta) \\ + P(R_i^{(t)} = 1, H_i^{(t)} = 0 \mid m_i^{(t)}, d_i, \Theta). \end{aligned} \tag{2}$$

Here we make our second assumption: reports cannot be false since the Bureau verifies reports prior to their addition to the SSA and we perform our own quality control. Hence the last term in Equation 2 is zero. The definition of conditional probability then yields

$$P(R_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta) = P(R_i^{(t)} = 1 \mid H_i^{(t)} = 1, m_i^{(t)}, d_i, \Theta) P(H_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta)$$

Further we assumed that knowledge of severe hail, by conditioning upon $H_i^{(t)} = 1$, encapsulates information in $m_i^{(t)}$. Whilst larger hail is generally more likely to be reported (Williams et al., 1999; Cintineo et al., 2012; Kelly et al., 1985), MESH has limited skill in inferring hail size (Witt et al., 1998; Warren et al., 2020; Cintineo et al., 2012; Brimelow and Taylor, 2017; Murillo and Homeyer, 2019) and so we do not employ it for this purpose in our reporting model. Moreover, we assumed that hailstorms are ambivalent to human presence and so the density measurement does not inform the actual probability of hail fall. These assumptions leave us with

$$P(R_i^{(t)} = 1 \mid m_i^{(t)}, d_i, \Theta) = P(R_i^{(t)} = 1 \mid H_i^{(t)} = 1, d_i, \Theta) P(H_i^{(t)} = 1 \mid m_i^{(t)}, \Theta). \tag{3}$$

Thus our model is composed of a *reporting model* and a *hail model* represented by each of the two factors in Equation 3 respectively. Whilst our primary interest is in the hail model which links MESH to probabilities of hail, as from this model we can infer hail probabilities from a MESH time series, we need also describe the reporting process to linking hail to the reports we observe. The way in which we construct these models is governed by the parameter Θ. Whilst we have constructed Equation 3 for our specific application, we note that these models could be far mare general in nature: $m_i^{(t)}$ and d_i could each be replaced with several covariates as desired and the hail variable too could be easily changed to other binary indicators of interest. Even our assumption that all reports are legitimate could be relaxed so long as we then modelled the process governing false reports as a function of the available covariates.

When constructing a Bayesian model we require both a model for the data given the parameters, the likelihood, and the priors on these parameters (Gelman et al., 2004). To expand Equation 3 into a likelihood function for the full data, we assumed further that, conditional upon $m_i^{(t)}$ and d_i the probability of hail in a any grid cell is independent of any other. This assumption is mathematically convenient, but also stringent. Notably, under this assumption we fail to capture the possibility of hail advection between cells. Whilst we still obtained acceptable results, it could be relaxed by incorporating additional predictors into the hail model, like the MESH values of neighbouring cells. Finally, we must specify

forms for the reporting and hail models to fully specify the likelihood. We employ logistic functions for each, namely

$$P(R_i^{(t)} = 1 \mid H_i^{(t)} = 1, d_i, \Theta) = \frac{1}{1 + \exp\left[-(\delta_1 f\left(d_i\right) + \delta_2)\right]} \tag{4}$$

and

$$P(H_i^{(t)} = 1 \mid m_i^{(t)}, \Theta) = \frac{1}{1 + \exp\left[-(\mu_1 g\left(m_i^{(t)}\right) + \mu_2)\right]} \tag{5}$$

where

$$f\left(d\right) = \log d$$

and

$$g\left(m\right) = \log\left(\left(\frac{m}{2.54}\right)^2 + 1\right)$$

reduces the skew in the MESH values (by first transforming MESH to the significant hail index, see Witt et al. (1998)) and $\Theta = (\delta_1, \delta_2, \mu_1, \mu_2)$. The logistic functions could be replaced by other parametrisations, lending generality to our model. A caveat to this flexibility is the importance of selecting functions that are non-linear in the unknown parameters. If not, it becomes impossible to separate the influence on a grid cell's reporting probability of each of the two factors as they can be scaled arbitrarily.

Lastly, we must specify priors. We found that, of the possibilities we tested, weakly informative priors upon each of the four parameters best enabled the data to drive the posterior whilst minimising bias. Nonetheless, we strongly suspect that reports are more likely over densely populated areas and severe hail more likely at higher MESH values. Hence we enforced that Equation 5 and Equation 4 be monotonically increasing by requiring $d_1 > 0$ and $m_1 > 0$. Hence our priors are

$$\begin{aligned}
\delta_1 &\sim \log\mathcal{N}\left(-\log 2, \left(\frac{1}{2}\right)^2\right), \\
\delta_2 &\sim \mathcal{N}\left(-5, 2^2\right), \\
\mu_1 &\sim \log\mathcal{N}\left(0, 1^2\right), \text{ and} \\
\mu_2 &\sim \mathcal{N}\left(-3, 1^2\right)
\end{aligned}$$

where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution with mean μ and variance σ^2 and $\log\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the log-normal distribution with the same parameters.

C Idealised Parameter Confounding Experiments

In these simulations, the observed MESH time series for each grid cell and the associated population density remained unchanged. However, the probabilities in Equation 4 and Equation 5 were calculated using known parameters, specified in Table 1. Reports were then simulated from independent Bernoulli distributions governed by these probabilities and hence data was generated in the exact way specified by our model.

In Table 1 we use the notation $\mathcal{N}\left(\mu, \sigma^2\right)$ to denote the normal distribution with mean μ and variance σ^2. The notation $\mathcal{N}_{\text{trunc}}\left(\mu, \sigma^2\right)$ indicates the normal distribution was truncated with a lower bound of zero.

D Reliability Diagram

We constructed the reliability diagram following Bröcker and Smith (2007). The 95% bootstrap confidence intervals are derived from 1000 simulations supposing the expected probabilities from the model posterior distribution were perfectly reliable. Thus, the uncertainty bounds represent the innate variability within each interval of a reliable forecast. For this reason, there is greater uncertainty when where are fewer observations represented by the data point.

Table 1: Parameters and prior distributions used in the idealised simulations.

(a) Hail probability parameters.

		m_1		m_2	
Regime	Priors	Value	Prior	Value	Prior
LHHR	Strongly informative	1	$\mathcal{N}_{\text{trunc}}\left(1, 1^2\right)$	-6.9	$\mathcal{N}\left(-6.9, 1^2\right)$
	Weakly informative		$\mathcal{N}_{\text{trunc}}\left(1, 3^2\right)$		$\mathcal{N}\left(-6.9, 3^2\right)$
	Misleading		$\mathcal{N}_{\text{trunc}}\left(3, 1^2\right)$		$\mathcal{N}\left(-8, 1^2\right)$
HHLR	Strongly informative	3	$\mathcal{N}_{\text{trunc}}\left(3, 1^2\right)$	-8.0	$\mathcal{N}\left(-8, 1^2\right)$
	Weakly informative		$\mathcal{N}_{\text{trunc}}\left(3, 3^2\right)$		$\mathcal{N}\left(-8, 3^2\right)$
	Misleading		$\mathcal{N}_{\text{trunc}}\left(1, 1^2\right)$		$\mathcal{N}\left(-6.9, 1^2\right)$

(b) Conditional reporting probabilty parameters.

		d_1		d_2	
Regime	Priors	Value	Prior	Value	Prior
LHHR	Strongly informative	1.0	$\mathcal{N}_{\text{trunc}}\left(1, 1^2\right)$	-2	$\mathcal{N}\left(-2, 1^2\right)$
	Weakly informative		$\mathcal{N}_{\text{trunc}}\left(1, 3^2\right)$		$\mathcal{N}\left(-2, 3^2\right)$
	Misleading		$\mathcal{N}_{\text{trunc}}\left(0.5, 1^2\right)$		$\mathcal{N}\left(-4, 1^2\right)$
HHLR	Strongly informative	0.5	$\mathcal{N}_{\text{trunc}}\left(0.5, 1^2\right)$	-4	$\mathcal{N}\left(-4, 1^2\right)$
	Weakly informative		$\mathcal{N}_{\text{trunc}}\left(0.5, 3^2\right)$		$\mathcal{N}\left(-4, 3^2\right)$
	Misleading		$\mathcal{N}_{\text{trunc}}\left(1, 1^2\right)$		$\mathcal{N}\left(-2, 1^2\right)$

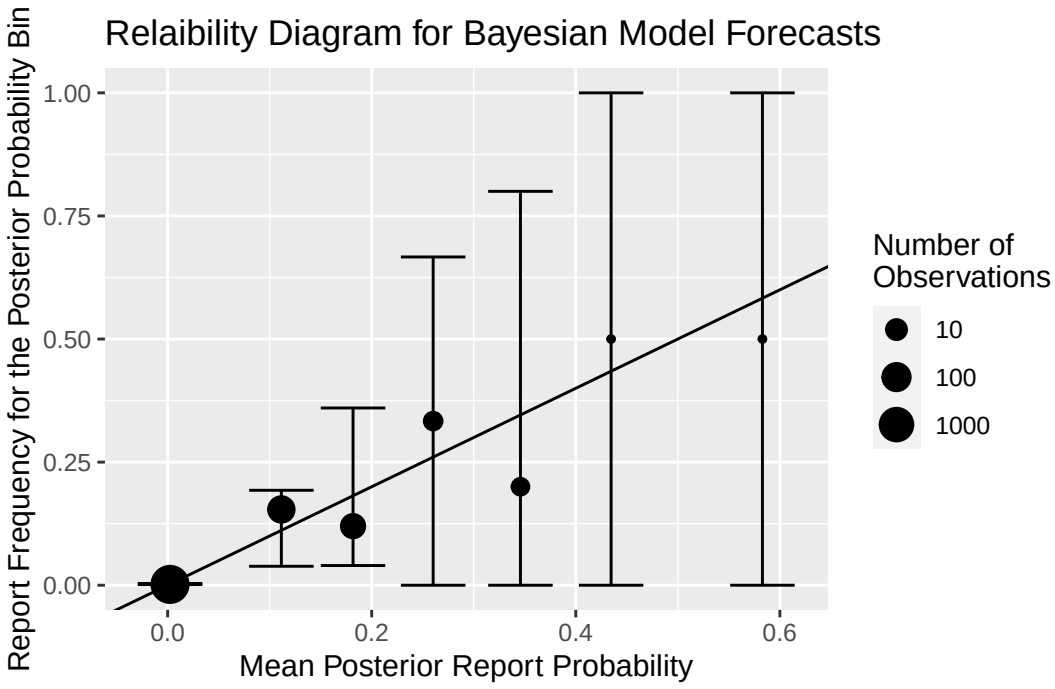

Figure 3: Reliability diagram for the test set (September 2015 - April 2016) storm season forecasts with an associated 95% confidence interval.