Rhoades, D. A., Christophersen, A., Gerstenberger, M. C., Liukis, M., Silva, F., Marzocchi, W., ... Jordan, T. H. (2018). Highlights from the first ten years of the New Zealand earthquake forecast testing center. *Seismological Research Letters*, *89*(4), 1229-1237. https://doi.org/10.1785/0220180032

HIGHLIGHTS FROM THE FIRST TEN YEARS OF THE NEW ZEALAND EARTHQUAKE FORECAST TESTING CENTER

David A. Rhoades[1], Annemarie Christophersen[1], Matthew C. Gerstenberger[1], Maria Liukis[2], Fabio Silva[3], Warner Marzocchi[4], Maximilian J. Werner[5] and Thomas H. Jordan[3]

[1]GNS Science, 1 Fairway Drive, Avalon, Lower Hutt 5010, New Zealand, d.rhoades@gns.cri.nz

[2]Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109

[3]Southern California Earthquake Center, University of Southern California, 3651 Trousdale Parkway, Los Angeles, California, 90089-0742

[4]Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605 - 00143 Rome, Italy

[5]University of Bristol, Tyndall Avenue, Bristol, BS8 1TH, United Kingdom

ABSTRACT

We present highlights from the first decade of operation of the New Zealand Earthquake Forecast Testing Centre of the Collaboratory for the Study of Earthquake Predictability (CSEP). Most results are based on reprocessing using the best available catalog, because the testing center did not consistently capture the complete real-time catalog. Tests of models with daily updating show that aftershock models incorporating Omori-Utsu decay can outperform long-term smoothed seismicity models with probability gains up to 1000 during major aftershock sequences. Tests of models with three-month updating show that several versions of the EEPAS (Every Earthquake a Precursor According to Scale) model, incorporating the precursory scale increase phenomenon and without Omori-Utsu decay, and the Double-Branching model, with both Omori-Utsu and exponential decay in time, outperformed a regularly updated smoothed seismicity model. In tests of five-year models over ten years without updating, a smoothed seismicity model outperformed the earthquake source model of the New Zealand national seismic hazard model. The performance of three-month and five-year models was strongly affected by the Canterbury earthquake sequence, which occurred in a region of previously low seismicity. Smoothed seismicity models were shown to perform better with more frequent updating. CSEP models were a useful resource for the development of hybrid time-varying models for practical forecasting after major earthquakes in the Canterbury and Kaikōura regions.

INTRODUCTION

The New Zealand Earthquake Forecast Testing Center (Gerstenberger & Rhoades, 2010) is a computer system established in 2008 as one of the regional testing centers of the Collaboratory for the Study of Earthquake Predictability (CSEP; Zechar et al.2010). Its purpose is to transparently test earthquake forecasting models against future earthquakes in

the testing region in defined magnitude classes, i.e., target earthquakes. Several major earthquakes have occurred in the testing region (Figure 1) during the past decade, including the 2010 M7.1 Darfield earthquake and its aftershocks (the Canterbury earthquakes) and the 2016 M7.8 Kaikōura earthquake. The Canterbury earthquakes stimulated two important retrospective experiments (Rhoades et al. 2016; Cattania et al. 2018, this issue). Prospective testing is being conducted for four classes of models, grouped according to the time-period for updating: five-years (time-invariant), six-months, three-months and one-day. Here we summarise the performance of the five-year models over 10 years of testing, of the three-month models from their inception in 2009 to 2017, and of the 1-day models during the Darfield and Kaikōura aftershock sequences. We also compare information gains between models from different classes during these two sequences.

EARTHQUAKE CATALOG TRANSITION AND LATENCY

The GeoNet Catalog is the authoritative information source for earthquake data supplied to the testing center (Gerstenberger and Rhoades, 2010). During 2012 there was a transition in GeoNet processing of the earthquake catalog from the CalTech-USGS seismic processor (CUSP) system (Lee et al., 1989) to the more automatic SeisComP3 (SC3) system (Hanka et al., 2010). The quality of the near real-time earthquake catalog during the CUSP period was negatively affected by large backlogs in processing of major earthquake sequences. At the beginning of the SC3 period, GeoNet temporarily withdrew the first few months of the SC3 catalog for a time, after anomalies in earthquake locations were noticed. Earthquake magnitudes were also affected by the transition. Both CUSP and SC3 produce local magnitudes, but CUSP used attenuation relations adjusted to New Zealand while SC3 was installed with the Californian attenuation relations. Statistical comparisons show that magnitudes of small earthquakes are lower in the SC3 catalogue but for $M \geq 5$ the

magnitudes are similar to those in the CUSP catalog (Rhoades et al. 2015). The near real-time catalog has thus been of variable quality over the ten years of testing and has included sizeable gaps at certain times. Moreover, the testing center was correctly configured to upload the SC3 catalog only between September 2014 and August 2016. Consequently, there was no real-time processing of forecasts during the aftershock sequence of the M7.8 Kaikōura earthquake of 13 November 2016 (UTC).

The five-year and three-month model tests have now been reprocessed using the best catalog available at the end of 2017 – a composite of the CUSP catalog up to 2012 and the SC3 catalog from 2013 on. So far, there has only been time to reprocess the one-day models for selected periods. The one-day models have been reprocessed from 10 November 2016 (UTC) through 13 February 2017 (UTC), in order to evaluate their performance during the first three months of aftershocks of the M7.8 Kaikōura earthquake. In contrast, the testing center was functioning well during the first year of the Canterbury sequence, which began with the M7.1 Darfield earthquake of 3 September 2010 (UTC), albeit with an input catalog that was incomplete, especially following large earthquakes. For example, there were six $M \geq 3.95$ aftershocks within 24 hours of the Darfield earthquake in the real-time catalog. This increased to 20 when all aftershocks were processed about 18 months later (Christophersen et al. 2013)

FIVE-YEAR MODELS

Five-year models were supplied to the testing center as fixed forecasts, similar to the Regional Earthquake Likelihood Models (RELM) experiment in California (Schorlemmer and Gerstenberger, 2007). The target earthquakes are those with magnitude $M \geq 4.95$ and hypocentral depth $h \leq 40$ km. There are two 5-year model classes – for “all earthquakes” and

"mainshocks only", respectively. The latter set of target earthquakes is obtained by declustering the whole catalog for magnitude $M \geq 1.95$ using the Reasenberg (1985) algorithm with standard parameter settings (Gerstenberger and Rhoades 2010) and then restricting the declustered catalog to $M \geq 4.95$.

There are five models in the 5-year classes (Table 1): a gridded version of the New Zealand national seismic hazard model (NZHM; Stirling et al. 2002), including both fault sources and distributed seismicity, a smoothed seismicity model based on proximity to past earthquakes (PPE; Jackson and Kagan, 1999; Rhoades and Evison, 2004) fitted to data up to the end of 2006, a version of PPE to be assessed against mainshocks only (PPE_DEC), a spatially uniform Poisson model (SUP) fitted to the test region up to the end of 2006 and included as a model of least information, and a version of SUP to be assessed against mainshocks only (SUP_DEC). The NZHM model was designed to forecast mainshocks only, but for reference purposes we include it in both classes. Spatial plots of models in the 5-year classes are given in Figure S1 of the electronic supplement. We analyse the performance of the models from 2008 Jan 1 to 2017 Dec 31, using the target events with magnitudes $M \geq 4.95$ from the finalised CUSP catalog and the SC3 catalog to 2017. The complete target catalog has 235 events and the declustered catalog has 49 "mainshocks" (Figure 1). The declustered catalog is listed in the electronic supplement (Table S1).

In the "all earthquakes" class, all models grossly under-predicted the number of target earthquakes in the test period, which had a much higher level of seismicity than in any period since at least the 1950s. Consistency of the models with the total number of target earthquakes is measured in CSEP by the N-test (Schorlemmer et al., 2007). All models failed the N-test with Poisson 95% confidence limits by a wide margin (Figure 2a), with vanishingly small p-values. This result is as much a reflection of the inadequacy of the

Poisson assumption as of the models themselves. The information gain of one model over another, taking into account the relative expected numbers of the models in spatial cells and magnitude bins (and time periods, if updating is involved) to which the target earthquakes belong, is measured in CSEP by the T-test (Rhoades et al. 2011). According to the T-test, the PPE model performed significantly better than SUP with an information gain (per earthquake) of 0.22 ± 0.12 (95% confidence limits). The information gain (IG) is related to the probability gain (PG) by IG = ln(PG). Therefore, for this example, the probability gain is only 1.24. The SUP model in turn outperformed the NZHM model with an information gain of 0.35±0.11 (Figure 2b). The latter result may be attributed partly to the low expected number of earthquakes associated with NZHM as a mainshocks-only model and partly to the effect of the Canterbury earthquakes.

The Canterbury earthquakes, which contributed 54 earthquakes to the target set of all events with M ≥4.95 up to 2012, occurred in a region of previously low seismicity, not close to modelled faults with high slip rates. When the 10-year test period is split into two five-year periods, the information gain of SUP over NZHM is higher in the first period (2008-2012), which included the Canterbury earthquakes, than in the second (2013-2017), about 0.7 versus -0.05 (Figure 3).

In the "mainshocks only" class, all models passed the N-test at the Poisson 95% significance level (Figure 4a). The T-test shows that PPE_DEC was the best performing model, with an information gain of about 0.4 over NZHM and about 0.75 over SUP_DEC (Figure 4b).

THREE-MONTH MODELS

Processing of three-month models commenced in May 2009, with the first quarterly forecasts being produced in July 2009. The target earthquakes are all events with magnitude $M \geq 4.95$ and hypocentral depth $h \leq 40$ km. Because this class is designed for medium-term and not short-term models, the models in this class were provided with an input catalog ($M \geq 2.95$) that terminated one month before the start of each test period. The three-month models (Table 2) include five versions of the "Every Earthquake a Precursor According to Scale" (EEPAS) model (Rhoades and Evison, 2004, 2005, 2006) based on the precursory scale increase phenomonen (Evison and Rhoades, 2002, 2004), the PPE model updated at each forecast period, and the "Double Branching Model" (DBM; Marzocchi and Lombardi, 2008). For reference purposes, we also include the SUP model, scaled down from the five-year model, in this class.

Four versions of the EEPAS model as originally proposed are included. These are denoted EEPAS-0R, EEPAS-0F, EEPAS-1R and EEPAS-1F, where "0" indicates equal weighting of all earthquakes in the input catalog, "1" indicates down-weighting of aftershocks, "R" indicates a restricted set of only four parameters were fitted and "F" indicates a fuller set of six parameters were fitted. The experiment was thus an opportunity to evaluate the relative worth of two different weighting strategies and two different fitting strategies (Rhoades, Gerstenberger et al., 2008). Additionally, an earthquake-rate dependent EEPAS-0F model (ERDEEP) is included, in which the scaling parameters for precursor time and area depend on the local seismicity rate, as estimated by the PPE model. This variant of the EEPAS model was described by Rhoades et al. (2010).

The three-month models all significantly under-estimate the number of target earthquakes in the total test interval (Figure 5a). This underprediction is consistent with the fact that a high

proportion of the target earthquakes occurred as aftershocks of larger events. Except for the DBM model, these models make no attempt to forecast the Omori-Utsu decay of aftershocks and, with the gap between the end of the input catalog and the start of each test period, have little opportunity do so anyway.

The T-test shows EEPAS-0F to be the best-performing model in the total test interval (Figure 5b). The same model was the best performing model in a similar analysis of 3-month models in the CSEP California testing center (Schneider et al., 2014). Again, there are differences in relative model performance within the test period. Up to the end of 2012, DBM was marginally the most informative model and none of the EEPAS models significantly outperformed PPE (Figure 6a). However, from 2013 on, all of the EEPAS models significantly outperformed DBM and PPE (Figure 6b) with information gains over PPE of about 0.6. Again, the results up to the 2012 are strongly affected by the Canterbury earthquake sequence. Numerous large aftershocks more than three months after the initiating Darfield earthquake provided an opportunity, from the beginning of 2011 on for the DBM model, which includes Omori-Utsu aftershock decay, to perform well, and for the PPE model to incorporate the early Darfield aftershocks. In contrast, the EEPAS models benefited less from the early aftershocks, because their response to any new earthquake begins only gradually and peaks several months or years later, depending on its magnitude. Selected spatial plots of the 3-month models are given in Figures S2-S4 of the electronic supplement.

ONE-DAY-MODELS

Major aftershock sequences that occurred during the past decade provide an opportunity to examine the performance of the one-day models in the conditions under which aftershock models are expected to perform best. The target earthquakes for one-day testing are the

events with $M \geq 3.95$ and $h \leq 40$ km. The one-day models (Table 3) include the Short-Term Earthquake Probabilities (STEP) model (Gerstenberger et al. 2004, 2005), a variation of STEP incorporating a revised estimation of aftershock abundance (STEP_ABU; Christophersen and Gerstenberger, 2010), and an Epidemic Type Aftershock (ETAS) model (Ogata, 1988, 1998). The installed version of ETAS was described by Rhoades (2013) and Rhoades, Gerstenberger et al. (2008). A version of PPE with daily updating (PPE_1d) is also installed as a reference model with no Omori-type aftershock behaviour. We present results for two periods of intense aftershock activity: six months starting the first day after the M7.1 Darfield earthquake of 2010 Sep 3 (the Canterbury earthquakes), and three months starting the first day after the M7.8 Kaikōura earthquake of 2016 Nov 13. The Canterbury results were obtained in near real-time testing, with a one-month delay in processing; the Kaikōura results are based on the best catalog available in December 2017. No results for the STEP or STEP_ABU models are presented in the Canterbury results, because of a problem with the installation of these models at that time. Irregularities still affected these models at the time of the Kaikōura aftershocks. First, the forecasts were computed only for magnitudes $M \geq 4.95$. However, since these models follow the Gutenberg-Richter relation (Gutenberg and Richter, 1944) in each spatial cell, this has been corrected by post-processing to extend the forecasts down to $M \geq 3.95$. Secondly, the background has lower rates than NZHM, which is the intended background model. (c.f. Figures S1, S5 and S6 of the electronic supplement). We calculate and test a modified model, STEP_MOD, in which the cells with lower rates than NZHM are replaced by the rates of the NZHM model, to measure the effect of the low background on the performance.

The probability gain of aftershock models over long-term smoothed seismicity models is of interest. Therefore, we have included as additional reference models in the one-day tests the

stationary five-year PPE model and the PPE model with 3month updating, with expected numbers scaled down to one-day (PPE_5y and PPE_3m, respectively). For the Kaikōura tests, these models, which conform to the Gutenberg-Richter relation, have also been extended down to $M \geq 3.95$.

In the case of Canterbury, the ETAS model passed the N-test. The PPE reference models under-estimated the number of target events, as expected (Figure 7a). The information gain of the ETAS model is about 4.5 relative to the PPE_1d model (Figure 7b) and almost 7 relative to PPE_5y. The latter value corresponds to a probability gain of approaching 1000. In the case of Kaikōura, the forecasts of STEP and STEP_ABU were identical, and therefore only the STEP model is presented. The STEP and STEP_MOD models overestimated the number of target earthquakes by a factor of about three and the ETAS model by about 20 percent (Figure 8a). The information gain of the ETAS model is about 4 relative to the PPE-1d model and about 6 relative to the PPE_5y model (Figure 8b). The latter value corresponds to a probability gain of about 400. A lower information gain for Kaikōura than for Canterbury is to be expected, because the Canterbury earthquakes occurred in a region of previously low seismicity and the Kaikōura earthquake and its aftershocks occurred in a region of previously high seismicity. Figure 8 shows that there is very little difference in the performance of the STEP and STEP_MOD models, because of the small number of target events in the background area during the period analysed.

DISCUSSION

The inconsistent results of the five-year and three-month models in different sub-periods shows that results obtained for one period, even when supported by many target events, will not necessarily be repeated in other periods, owing to the non-stationarity of the earthquake-

generating process. As with any CSEP test, the results are specific to the time period examined and can be considered only indicative of how the models might perform in other time periods. For this reason, care must be taken in interpreting the results and greater confidence can be placed in model comparisons that are confirmed by consistent results over multiple time-periods and multiple regions.

Recent research helps to explain the relatively poor performance of the EEPAS models during the Canterbury earthquakes. Unlike most major earthquakes, the Darfield earthquake does not have a precursory scale increase in the instrumental catalog. A study using the physics-based earthquake simulator RSQSim indicates that, given the low strain rate in the Canterbury area, a precursor time much longer than the existing instrumental catalog would be expected for the Darfield earthquake (Christophersen et al. 2017). The present formulation of the EEPAS model does not allow for this effect of low strain rate.

Better long-term models than those currently installed in the testing center as five-year models have been developed and tested retrospectively during the past decade, but not yet installed in the CSEP testing center. These are hybrid models that combine data on past earthquake occurrence, fault locations with associated slip rates, and strain rates. Rhoades and Stirling (2011) showed that an additive mixture of PPE with an earthquake likelihood model (PMF), based on proximity to mapped faults weighted by slip rate, produced an information gain over PPE of about 0.1 over a test period from 1997-2006. Rhoades et al. (2015) showed that a multiplicative hybrid involving multiple fault and earthquake covariates had an information gain of 0.05 – 0.2 over PPE, with somewhat smaller gains when the PPE model is updated to the start of the test period. Rhoades et al. (2017) showed that when shear

strain is added to the pool of covariates contributing to a multiplicative hybrid, the information gain is further increased by about 0.3 for the 2012-2015 test period.

The performance of the various versions of PPE as reference models for one-day testing shows that smoothed seismicity models perform much better when recently updated. For example, the information gain of the three-month PPE model over SUP was 0.75 (Figure 5b), larger than that of the five-year PPE model over SUP, which was 0.22 (Figure 6b); and the information gains of the PPE_1d over PPE_3m during the Darfield and Kaikōura aftershock sequences were 1.2 and 0.8 respectively (Figures 7b and 8b). The dependence of performance on recent updating is a natural consequence of earthquake clustering on a range of timescales. It is a factor to be considered when developing earthquake source models for long-term seismic hazard and a risk studies.

The information gain of PPE over NZHM is somewhat surprising, given that NZHM includes information on both faults and earthquakes. However, the two models used different selections of data from the earthquake catalog and different smoothing methods. Also, NZHM was designed with time-periods of many decades in mind; the tests conducted here do not show how it would perform over very long periods. The information gain of SUP over the NZHM in the all-earthquakes class can be attributed mainly to the fact that NZHM was designed to forecast mainshocks only. Nevertheless, the occurrence of the disastrous M6.2 Christchurch earthquake of 2011 Feb 22, the largest aftershock so far of the Darfield earthquake, made it clear that a time-invariant model that aims to forecast mainshocks only is not likely to perform well in estimating seismic hazard over the next few decades in Canterbury. Consequently, in the wake of the Christchurch earthquake, there was a demand for a new seismic hazard model for Canterbury for the next 50 years. The earthquake source

model had to allow for time-varying earthquake occurrence and to cover timescales up to 50 years, to support decision-making for the recovery of Christchurch (Gerstenberger et al. 2014).

CSEP models provided a valuable resource for the construction of the new Canterbury source model. The models already installed in the testing center became useful building blocks for a hybrid model for forecasts on all required timescales. A time-varying component was defined as a mixture of time-varying models (STEP, ETAS and versions of EEPAS), and a long-term component was defined as a mixture of smoothed seismicity models with a variety of data selections and smoothing methods. The mixture weights were assessed by expert elicitation. The hybrid was defined as the maximum of the time-varying and long-term components in each spatial cell and magnitude bin, as in the definition of the STEP model (Gerstenberger, 2005; Gerstenberger et al., 2014, 2016).

Testing of the Canterbury earthquake source model was undertaken in the New Zealand Testing Center. A retrospective experiment was carried out in which the hybrid model was compared with component models in a series of one-year forecasts with lags up to 25 years. The experiment confirmed that hybrid model outperformed all, or nearly all, of its components at all time-lags (Rhoades et al. 2016).

A 100-year hazard model for central New Zealand was developed following the Kaikōura earthquake for planning of road and rail reconstruction, using a hybrid source model with a similar form, but with two differences: the hybrid is defined as the maximum of three components (short-term, medium-term and long-term), and the long-term component includes a contribution from a multiplicative hybrid incorporating strain rates.

Studies of hybrid models in New Zealand and elsewhere (Rhoades and Gerstenberger, 2009; Rhoades and Stirling, 2012; Marzocchi et al. 2012; Taroni et al. 2013; Rhoades, 2013; Steacy et al. 2014; Rhoades et al., 2014, 2015, 2016, 2017) indicate that hybrid forecasting models can usually outperform individual models that are based on restricted assumptions and data. CSEP's initial emphasis has been on testing the consistency and information value of individual models. However, for practical forecasting, it may be more useful to test whether a new model or data stream can be combined with existing models into a more informative hybrid model. This suggests that CSEP should extend its formal testing procedures to hybrid models.

CONCLUSION

The occurrence of an unusually large number of earthquakes in New Zealand over the past decade has been helpful for testing of CSEP models, especially in the one-day class. One-day model testing showed that the ETAS model gave probability gains of 400 to 1000 over a long-term PPE model during the Kaikōura and Darfield Aftershock sequences. Three-month model testing showed that the best-performing EEPAS model gave an information gain of 0.5 over PPE. The PPE smoothed seismicity model outperformed the NZHM model over ten years of testing. Concurrent research suggests that long-term earthquake source models can be improved by combining earthquake and fault data in different ways than they have traditionally been combined, and by the incorporation of strain rates. The performance of smoothed seismicity models is strongly improved by regular updating. CSEP models were a valuable resource for the development of practical time-varying hybrid forecasting models in the wake of the Canterbury and Kaikōura earthquakes. Tests of hybrid models inside and outside the testing center indicate that hybrid models usually outperform simpler models.

How best to construct and test hybrid forecasting models is an important problem for CSEP in the future.

DATA AND RESOURCES

The New Zealand Earthquake Forecast Testing Center depends on data provided by GeoNet at http://www.geonet.org.nz, and in particular on the GeoNet earthquake catalog at http://wfs.geonet.org.nz, last accessed January 2018.

ACKNOWLEDGEMENTS

This research was supported by the New Zealand Strategic Science Investment Fund, the New Zealand Earthquake Commission and the Southern California Earthquake Center (Contribution No. 8012). SCEC is funded by NSF Cooperative Agreement EAR-1033462 & USGS Cooperative Agreement G12AC20038. Helpful internal reviews of the manuscript were given by C. Mueller and R. Buxton. Constructive reviews were also provided by R. Console, an anonymous reviewer and Guest Editor A. Michael.

REFERENCES

Christophersen, A., and M.C. Gerstenberger (2010). A new generic model for aftershock occurrence. Appendix I in Rhoades, D.A., M.C. Gerstenberger, and A. Christophersen. Development, installation and testing of new models in the New Zealand Earthquake Forecast Testing Centre. *GNS Science Consultancy Report CR 2010/253*, 25 pp.

Christophersen, A., D. A. Rhoades, S. Hainzl, E. G. C. Smith, and M. C. Gerstenberger (2013), The Canterbury sequence in the context of global earthquake statistics. *GNS Science Consultancy Report CR 2013/196*, 28pp.

Christophersen, A., D.A. Rhoades, and H.V. Colella (2017). Precursory seismicity in regions of low strain rate: insights from a physics-based earthquake simulator. *Geophys. J. Int.* **209**, no.3, 1513-1525, doi: 10.1093/gji/ggx104

Evison, F.F., and D.A. Rhoades (2004). Demarcation and Scaling of Long-term Seismogenesis, *Pure Appl. Geophys.* **161**, no.1, 21-45.

Gerstenberger M., S. Wiemer, and L. Jones (2004). Real-time forecasts of Tomorrow's Earthquakes in California: a New Mapping Tool, *United States Geological Survey Open-File Report, 2004-1390*.

Gerstenberger, M.C., S. Wiemer, L.M. Jones, and P.A. Reasenberg (2005). Real time forecasts of tomorrow's earthquakes in California, *Nature* **435**, 328-331.

Gerstenberger, M.C., and D.A. Rhoades (2010). New Zealand Earthquake Forecast Testing Centre. *Pure Appl. Geophys.* **167,** no. 8/9, 877-892.

Gerstenberger, M.C., G.H. McVerry, D.A. Rhoades and M.W. Stirling (2014). Seismic hazard modeling for the recovery of Christchurch, New Zealand. *Earthquake Spectra,* **30,** no. 1, 17-29, doi: 10.1193/021913EQS037M.

Gerstenberger, M.C., D.A. Rhoades, and G.H. McVerry (2016). A hybrid time-dependent probabilistic seismic-hazard model for Canterbury, New Zealand. *Seismol. Res. Lett.*, **87**, no. 6, 131--1318; doi: 10.1785/0220160084.

Gutenberg, B., and C.F. Richter (1944). Frequency of earthquakes in California. *Bull. Seismol. Soc. Am.* **34**, 185-188.

Hanka, W., J. Saul, B. Weber, J. Becker, P. Harjadi, Fauzi, and GITEWS Seismology Group (2010). Real-time earthquake monitoring for tsunami warning in the Indian Ocean and

beyond, *Nat. Hazards Earth Syst. Sci*., **10**, 2611-2622, doi:10.5194/nhess-10-2611-2010.

Jackson, D.D., and Y.Y. Kagan (1999). Testable Earthquake Forecasts for 1999, *Seismol. Res. Lett.* **70**, no.4, 393-403.

Lee, W.H.K., and S.W. Stewart (1989). Large-scale processing and analysis of digital waveform data from the USGS Central California microearthquake network, In: Litehiser, J.J. (Ed.), Observatory Seismology: An Anniversary Symposium on the Occasion of the Centennial of the University of California at Berkeley Seismographic Stations. University of California Press. p. 86.

Marzocchi, W., and A.M. Lombardi (2008). A double branching model for earthquake occurrence, *J. Geophys. Res.* **113**, no. B8, B08317.

Marzocchi, W., J.D. Zechar, and T.H Jordan (2012). Bayesian forecast evaluation and ensemble earthquake forecasting. *Bull. Seismol. Soc. Am*,, **102**, no. 6, pp. 2574-2584, doi: 10.1785/0120110327

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Am. Stat. Ass*. **83**, 9-27.

Ogata, Y. (1998). Space-time point process models for earthquake occurrences. *Ann. Inst. Stat. Math.* **50**, 379-402.

Rhoades, D.A. (2013). Mixture models for improved earthquake forecasting with short-to-medium time horizons. *Bull. Seismol. Soc. Am.*, **103**, no.4, 2203-2215; doi: 10.1785/0120120233

Rhoades, D.A., and M.C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting. *Bull. Seismol. Soc. Am.*, **99**, *no. 2a*, 636-646; doi: 10.1785/0120080063

Rhoades, D.A., M.C. Gerstenberger, and A. Christophersen (2010). Development, installation and testing of new models in the New Zealand Earthquake Forecast Testing Centre. *GNS Science Consultancy Report CR 2010/253*, 55 pp.

Rhoades, D.A., M.C. Gerstenberger, A. Christophersen, J.D. Zechar, D. Schorlemmer, M.J. Werner and T.H. Jordan (2014). Regional earthquake likelihood models II : information gains of multiplicative hybrids. *Bull. Seismol. Soc. Am.*, **104,** no. 6, 3072-3083; doi: 10.1785/0120140035

Rhoades, D.A., A. Christophersen, and M.C. Gerstenberger (2015). Multiplicative earthquake likelihood models based on fault and earthquake data, *Bull. Seismol. Soc. Am*. **105**, no. 6, 2955-2968.

Rhoades, D.A., A. Christophersen, and M.C. Gerstenberger (2017). Multiplicative earthquake likelihood models incorporating strain rates, *Geophys. J. Int.* **208**, 1764–1774.

Rhoades, D.A., and F.F. Evison (2004). Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure Appl. Geophys.* **161**, no. 1, 47-72

Rhoades, D.A., and F.F. Evison (2005). Test of the EEPAS forecasting model on the Japan earthquake catalogue. *Pure and Applied Geophysics, 162(6/7):* 1271-1290

Rhoades, D.A., and F.F. Evison (2006). The EEPAS forecasting model and the probability of moderate-to-large earthquakes in central Japan. *Tectonophys.* **417**, no. 1/2, 119-130, doi: 10.1016/j.tecto.2005.05.051

Rhoades, D., M. Gerstenberger, A. Christophersen, M. Savage, and J. Zhuang, (2008). Testing and Development of Earthquake Forecasting Models, *GNS Science Consultancy Report 2008/70,* 107 pp.

Rhoades, D.A., M. Liukis, A. Christophersen, and M.C. Gerstenberger (2016). Retrospective tests of hybrid operational earthquake forecasting models for Canterbury. *Geophys. J. Int.*

204, no. 1, 440-456, doi: 10.1093/gji/ggv447.

Rhoades, D.A., D. Schorlemmer, M.C. Gerstenberger, A. Christophersen, J.D. Zechar, and M. Imoto (2011). Efficient testing of earthquake forecasting models. *Acta Geophysica* **59**, no. 4, 728-747; doi: 10.2478/s11600-011-0013-5

Rhoades, D.A., and M.W. Stirling (2012). An earthquake likelihood model based on proximity to mapped faults and cataloged earthquakes. *Bull. Seismol. Soc. Am.* **102**, no. 4, 1583-1599; doi: 10.1785/0120110326.

Schneider M, R. Clements, D.A. Rhoades, and D. Schorlemmer (2014). Likelihood- and residual-based evaluation of medium-term earthquake forecast models for California, *Geophys. J. Int.*, **198**, no. 3, 1307-1318.

Schorlemmer, D., and M.C. Gerstenberger (2007). RELM testing center, *Seismol. Res. Lett*. 78, no. 1, 30-36, doi: 10.1785/gssrl.78.1.30.

Schorlemmer, D., M.C. Gerstenberger, S. Wiemer, D.D. Jackson, D.A Rhoades (2007). Earthquake likelihood model testing, *Seismol. Res. Lett.* **78**, no.1 17-29.

Steacy, S., M.C. Gerstenberger, C.A.Williams, D.A. Rhoades, A. Christophersen (2014). A new hybrid Coulomb/statistical model for forecasting aftershock rates. *Geophys. J. Int*. **196**, no. 2 918-923, doi: 10.1093/gji/ggt404.

Stirling, M.W., G.H. McVerry,.and K.R. Berryman (2002) A new seismic hazard model for New Zealand, *Bull. Seismol. Soc. of Am.* **92**, 1878-1903

Taroni, M., J.D. Zechar, and W. Marzocchi (2013). Assessing annual global M6+ seismicity forecasts, *Geophys. J. Int*. **196**, no. 1, 422-431, doi: 10.1093/gji/ggt369.

Zechar, J.D., D. Schorlemmer, M. Liukis, J.Yu, F. Euchner, P.J. Maechling, T.H.Jordan (2010). The Collaboratory for the Study of Earthquake Predictability perspective on

computational earthquake science, *Concurrency Computation Practice and Experience*, **22**, no. 12, 1836-1847, doi: 10.1002/cpe.1519.

TABLES

Table 1: Overview of five-year models with no updating during the testing period.

Model	Target earthquakes	Description	Features	Reference
SUP	All	Stationary Uniform Poisson	Gutenberg-Richter (G-R) magnitude distribution	Rhoades & Evison, 2004
SUP_DEC	Mainshocks	As above	As above	As above
PPE, PPE_5y	All	Proximity to past earthquakes	Inverse power-law smoothing with magnitude weighting; G-R magnitude distribution	Jackson & Kagan, 1999; Rhoades & Evison, 2004
PPE_DEC	Mainshocks	As above	As above	As above
NZHM	Mainshocks	Earthquake source model of National Seismic Hazard Model	Characteristic earthquakes on faults and smoothed seismicity background	Stirling et al., 2002

Table 2: Overview of models with three-month updating

Model	Description	Features	Reference
DBM	Double-branching model	Omori-Utsu aftershock decay and longer-term exponential decay.	Marzocchi & Lombardi, 2008
EEPAS-0F	Every earthquake a precursor according to scale	Precursory scale increase (ψ) predictive scaling relations; equal weighting, eight fitted parameters	Evison & Rhoades, 2004; Rhoades & Evison, 2004, 2005, 2006
EEPAS-0R	As above	ψ predictive scaling relations; equal weighting, four fitted parameters	As above
EEPAS-1F	As above	ψ predictive scaling relations; aftershocks down-weighted, eight fitted parameters	As above
EEPAS-1R	As above	ψ predictive scaling relations; aftershocks down-weighted, four fitted parameters	As above
ERDEEP	Earthquake rate dependent EEPAS	ψ predictive scaling relations; equal weighting, eight fitted parameters	Rhoades et al., 2010
PPE, PPE_3m	Proximity to past earthquakes	See Table 1	See Table 1

Table 3: Overview of models with daily updating

Model	Description	Features	Reference
ETAS	Epidemic type aftershock sequence model	Omori-Utsu aftershock decay following every earthquake	Ogata, 1988, 1998; Rhoades, 2013
STEP	Short-term earthquake probability model	Superimposed Omori-Utsu aftershock decay sequences	Gerstenberger et al, 2004, 2005
STEP_ABU	STEP with adjustment for low numbers of aftershocks	as above	Christophersen & Gerstenberger, 2010
STEP_MOD	Modified STEP model	NZHM in background	See text
PPE_1d	Proximity to past earthquakes	See Table 1	See Table 1

CAPTIONS FOR FIGURES

Figure 1. Map of New Zealand showing test region (dotted polygon), search region for input catalog (dashed polygon) and epicenters of target earthquakes with $M \geq 4.95$ over the period 2008 Jan 1 to 2017 Dec 31 (235 events), classified as mainshocks (49 events) or aftershocks according to Reasenberg declustering.

Figure 2. Tests of five-year models targeting all earthquakes with $M \geq 4.95$, for the period 2008 Jan 1 to 2017 Dec 31. (a) N-tests comparing the actual number of target earthquakes (dashed line) with the expected number and its 95% confidence limits for each model under the Poisson assumption. (b) T-tests showing the information gain of other models and 95% confidence limits relative to the SUP model. The number of target earthquakes contributing to each comparison is also shown. For other T-test comparisons, see Table S2 of the electronic supplement.

Figure 3. T-tests of five-year models targeting all earthquakes with $M > 4.95$. (a) For the period 2008 Jan 1 to 2012 Dec 31, relative to NZHM; (b) For the period 2013 Jan 1 to 2017 Dec 31, relative to SUP. See caption of Figure 2 for more explanation. For other T-test comparisons, see Tables S3 and S4 of the electronic supplement.

Figure 4. Tests of five-year models targeting mainshocks only with $M \geq 4.95$, for the period 2008 Jan 1 to 2017 Dec 31. (a) N-tests. (b) T-tests relative to the NZHM model. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S5 of the electronic supplement.

Figure 5. Tests of three-year models targeting all earthquakes with M ≥ 4.95, for the period 2009 Jul 1 to 2017 Sep 30. (a) N-tests. (b) T-tests relative to the PPE model. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S6 of the electronic supplement.

Figure 6. T-tests of three-month models targeting all earthquakes with M ≥ 4.95, relative to PPE. (a) For the period 2009 Jul 1 to 2012 Dec 31; (b) For the period 2013 Jan 1 to 2017 Sep 30. See caption of Figure 2 for more explanation. For other T-test comparisons, see Tables S7 and S8 of the electronic supplement.

Figure 7. Tests of one-day models targeting earthquakes with M ≥ 3.95 and PPE models from the three-month and five-year classes (targeting earthquakes with M > 4.95) during the first six months of aftershocks of the Darfield earthquake from 2010 Sep 4 to 2012 Mar 8. (a) N-tests; (b) T-tests relative to PPE_1d. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S9 of the electronic supplement.

Figure 8. Tests of one-day models targeting all earthquakes with M ≥ 3.95 during the first three months of aftershocks of the Kaikōura earthquake, from 2016 Nov 14 to 2017 Feb 13. (a) N-tests; (b) T-tests relative to PPE_1d. See caption of Figure 2 for more explanation. For all T-test comparisons in this class, see Table S10 of the electronic supplement.

FIGURES

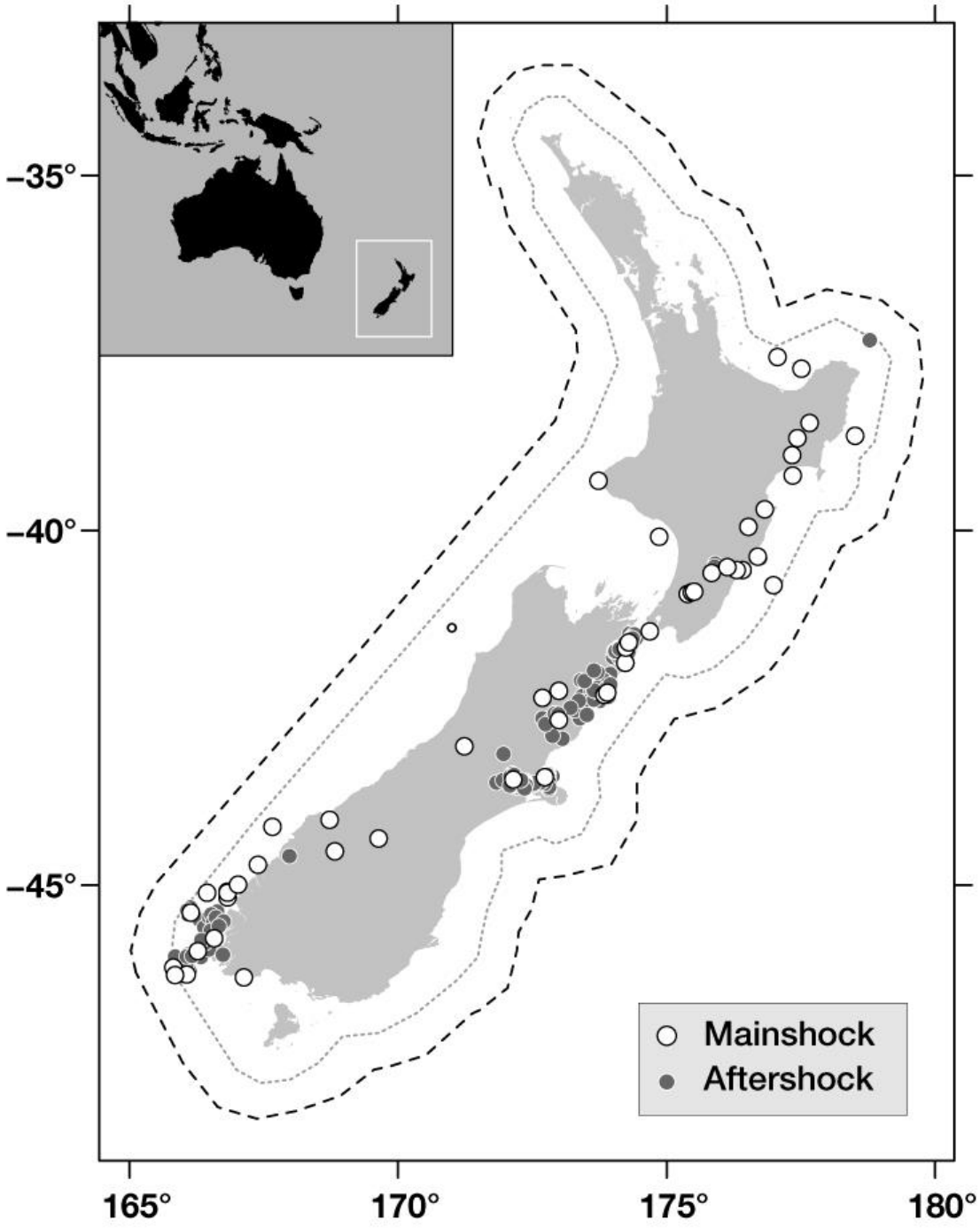

Figure 1. Map of New Zealand showing test region (dotted polygon), search region for input catalog (dashed polygon) and epicenters of target earthquakes with $M \geq 4.95$ over the period 2008 Jan 1 to 2017 Dec 31 (235 events), classified as mainshocks (49 events) or aftershocks according to Reasenberg declustering.

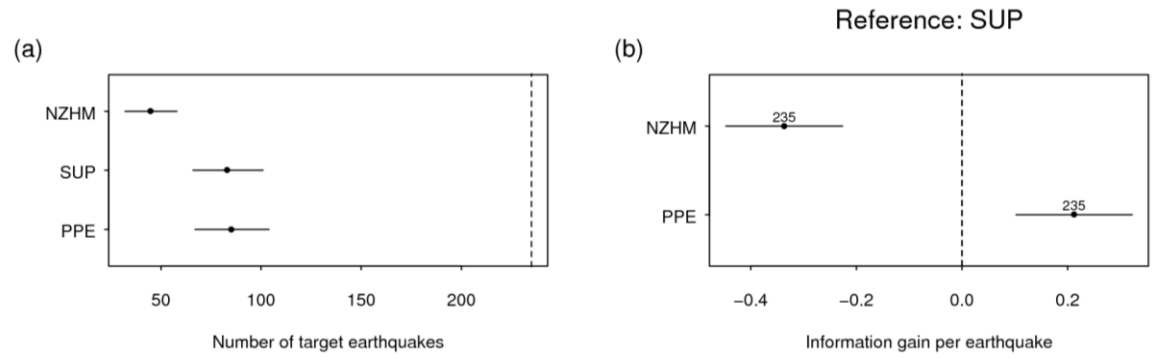

Figure 2. Tests of five-year models targeting all earthquakes with M ≥ 4.95, for the period 2008 Jan 1 to 2017 Dec 31. (a) N-tests comparing the actual number of target earthquakes (dashed line) with the expected number and its 95% confidence limits for each model under the Poisson assumption. (b) T-tests showing the information gain of other models and 95% confidence limits relative to the SUP model. The number of target earthquakes contributing to each comparison is also shown. For other T-test comparisons, see Table S2 of the electronic supplement.

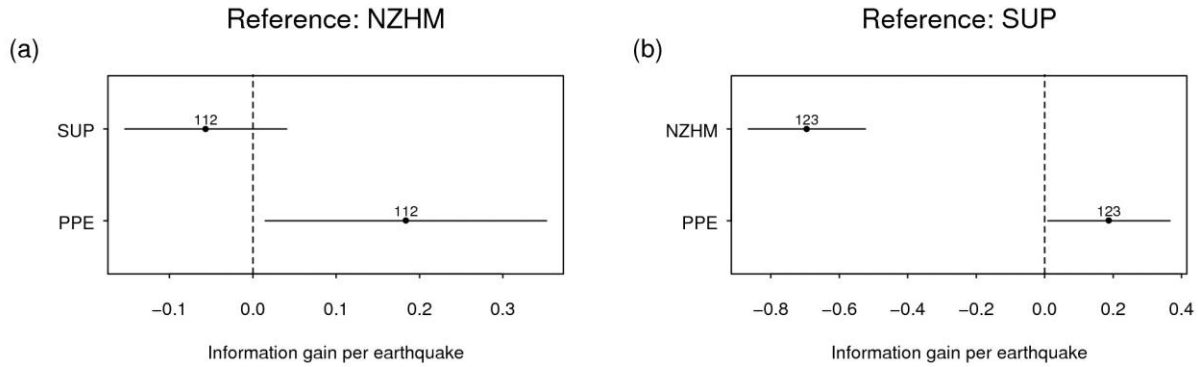

Figure 3. T-tests of five-year models targeting all earthquakes with M > 4.95. (a) For the period 2008 Jan 1 to 2012 Dec 31, relative to NZHM; (b) For the period 2013 Jan 1 to 2017 Dec 31, relative to SUP. See caption of Figure 2 for more explanation. For other T-test comparisons, see Tables S3 and S4 of the electronic supplement.

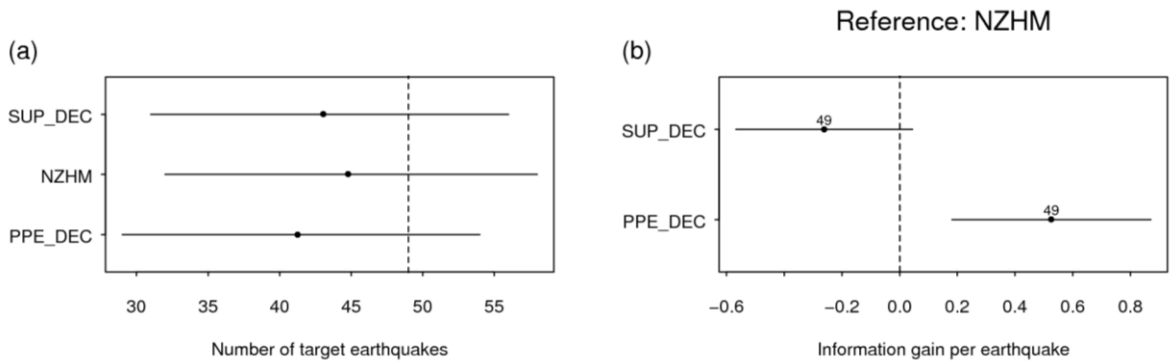

Figure 4. Tests of five-year models targeting mainshocks only with M ≥ 4.95, for the period 2008 Jan 1 to 2017 Dec 31. (a) N-tests. (b) T-tests relative to the NZHM model. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S5 of the electronic supplement.

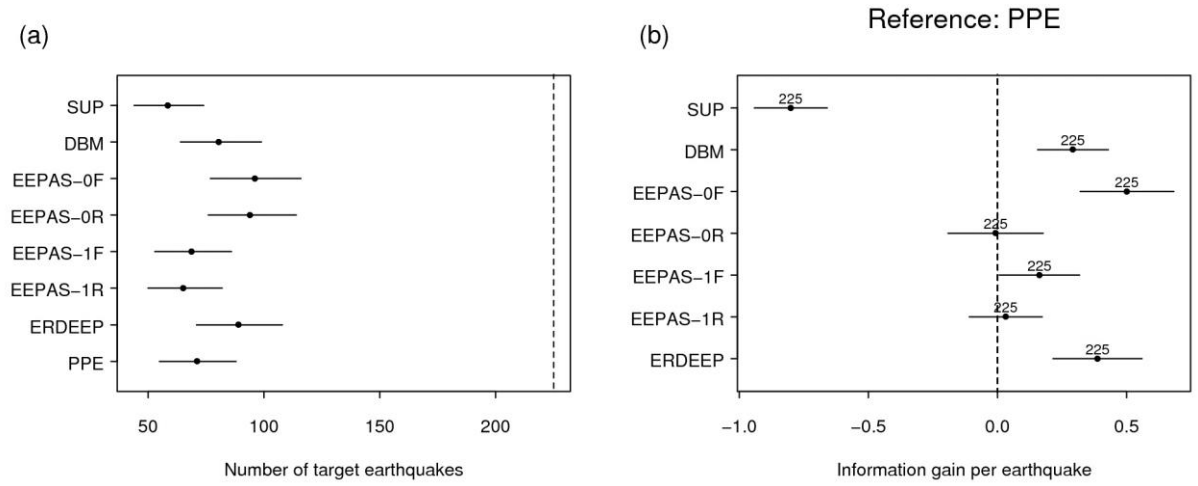

Figure 5. Tests of three-year models targeting all earthquakes with M ≥ 4.95, for the period 2009 Jul 1 to 2017 Sep 30. (a) N-tests. (b) T-tests relative to the PPE model. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S6 of the electronic supplement.

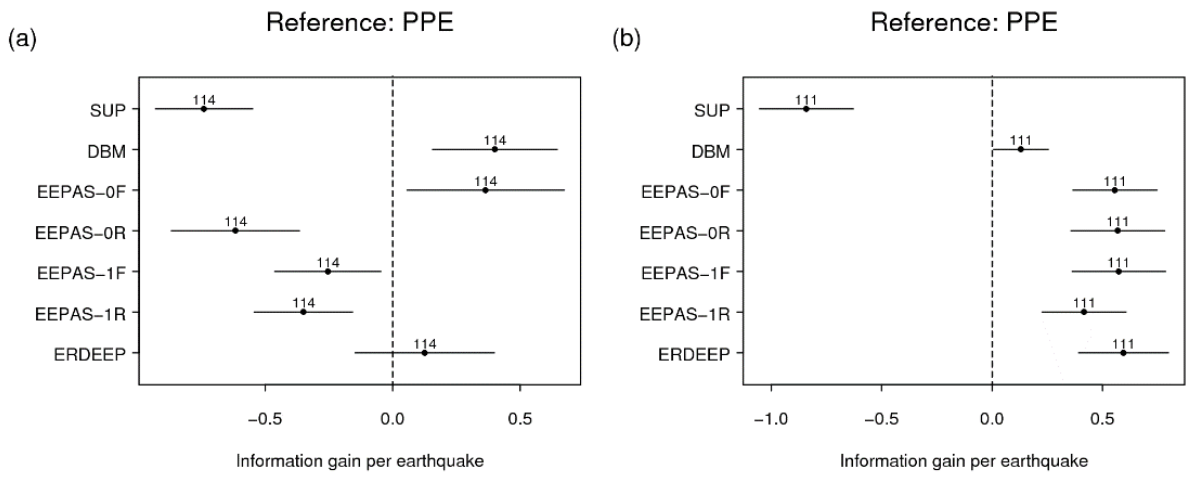

Figure 6. T-tests of three-month models targeting all earthquakes with M ≥ 4.95, relative to PPE. (a) For the period 2009 Jul 1 to 2012 Dec 31; (b) For the period 2013 Jan 1 to 2017 Sep 30. See caption of Figure 2 for more explanation. For other T-test comparisons, see Tables S7 and S8 of the electronic supplement.

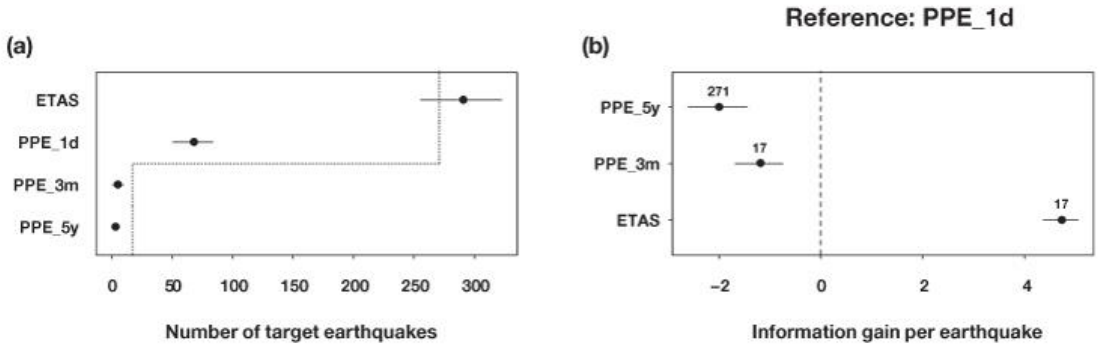

Figure 7. Tests of one-day models targeting earthquakes with M ≥ 3.95 and PPE models from the three-month and five-year classes (targeting earthquakes with M > 4.95) during the first six months of aftershocks of the Darfield earthquake from 2010 Sep 4 to 2012 Mar 8. (a) N-tests; (b) T-tests relative to PPE_1d. See caption of Figure 2 for more explanation. For other T-test comparisons, see Table S9 of the electronic supplement.

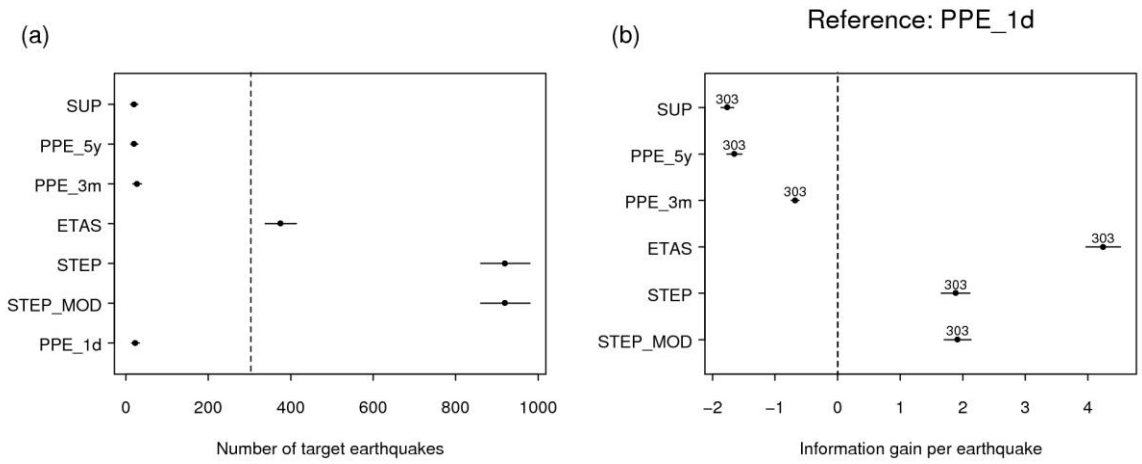

Figure 8. Tests of one-day models targeting all earthquakes with M ≥ 3.95 during the first three months of aftershocks of the Kaikōura earthquake, from 2016 Nov 14 to 2017 Feb 13. (a) N-tests; (b) T-tests relative to PPE_1d. See caption of Figure 2 for more explanation. For all T-test comparisons in this class, see Table S10 of the electronic supplement.