S&G 3753

GNS SCIENCE
TE PŪ AO

S+G
3753

Continued development of the New Zealand Earthquake Forecast Testing Centre

M. C. Gerstenberger
M. W. Stirling
A. Christophersen
D. A. Rhoades
R. Brownrigg

GNS Science Consultancy Report 2009/182
July 2009

DISCLAIMER

This report has been prepared by the Institute of Geological and Nuclear Sciences Limited (GNS Science) exclusively for and under contract to the Earthquake Commission. Unless otherwise agreed in writing by GNS Science, GNS Science accepts no responsibility for any use of, or reliance on any contents of this Report by any person other than the Earthquake Commission and shall not be liable to any person other than the Earthquake Commission, on any ground, for any loss, damage or expense arising from such use or reliance.

The data presented in this Report are available to GNS Science for other use from July 2009

BIBLIOGRAPHIC REFERENCE

Gerstenberger, M. C.; Rhoades, D. A.; Stirling, M.; Brownrigg, R. (VUW); Christophersen, A (ETH-Zurich). 2009. Continued development of the New Zealand Earthquake Forecast Testing Centre, *GNS Science Consultancy Report* CR 2009/182. 47p.

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

Recently the New Zealand Earthquake Forecast Testing Centre (NZTC) has been implemented within GNS Science. The NZTC is based on worked with the global Collaboratory for the Study of Earthquake Predictability (CSEP) and is: 1) a rigorous and community accepted rule-set for how earthquake forecasts, such as the National Seismic Hazard Model, should be evaluated; and 2) a computation environment that allows for transparent and reproducible tests of any forecast model implemented within the NZTC.

The NZTC allows for prospective and retrospective testing of forecast models and in this report we focus on the latter which test a forecast against historical earthquake data, including earthquake location and magnitude. We present the results of a suite of statistical tests that, collectively, can be used to understand the performance of a model against observed earthquakes. We have tested thirteen different models that are broken down into four classes which are based on the expected forecast length of each model: 1) one-day models; 2) three-month models; 3) six-month models; and 4) five-year models. The one-day model class is tested using observed earthquakes of magnitude 4.0 to 9.0 for all of New Zealand; all other models are tested using earthquakes of magnitudes 5.0 to 9.0.

The tests of the one-day models were hampered by an error in the CSEP testing procedure that was discovered during the testing, and were inconclusive. Five models were tested in the three-month category using data from 1996 to 2007, and all but one were shown to be consistent with the data. In a relative comparison test of the five models, the EEPAS-0F model was shown to provide a statistically significant improvement over the other four models. Only one model, M8 was tested in the six-month category and it was shown to be inconsistent with the earthquake data between 1996 to 2007. The five-year models were tested using observed earthquake data from 1984 to 2009 and all three models, a uniform Poisson model, a smoothed seismicity model, and the National Seismic Hazard Model, were shown to be consistent with the data. In a comparison test the smoothed seismicity model was able to significantly reject the other two models.

Also reported are two alternative testing routines to those used in the standard CSEP implementation. One improves the efficiency of an existing CSEP test to reduce unnecessary computation time and the other aims to provide a more powerful test of long-term forecasting models that aim to forecast for decades or more by using recorded ground motions and felt intensity reports.

Finally the M8 model was adapted to the CSEP requirements and appropriate computer code was written so that it could be evaluated in the testing centre.

TECHNICAL ABSTRACT

We have retrospectively tested thirteen earthquake forecast models within the New Zealand Earthquake Forecast Testing Centre (NZTC). Separating the models into four model classes (one-day, three-month, six-month, and five-year) we have evaluated the performance of the models using three likelihood-based tests which compare the model forecasts to observed earthquake data: the N-Test, which examines that total number of earthquakes forecast; 2) the L-Test which adds spatial and magnitude binning to the N-Test; and 3) the R-Test which, based on the L-Test, evaluates the relative performance of each model. In the L-Test and the R-Test the forecasts are evaluated using 0.1 degree cells and all classes are tested against observed data of magnitudes 5.0 to 9.0 except the one-day class which includes magnitudes as small as 4.0. The NZTC contains a regimented computational platform that is based on the work of the Collaboratory for the Study of Earthquake Predictability (CSEP) and uses a rule-set that was developed specifically for the New Zealand region (Gerstenberger, 2009).

The tests of the one-day models were hampered by an error in the CSEP testing procedure that was discovered during the testing, and the results were inconclusive. However, using testing outside of the NZTC environment, we were able to make model improvements. In this class we tested STEP (Gerstenberger, et al, 2005), ETAS (Rhoades, et al, 2008), Abundance (Christophersen, 2005) and a New STEP model (Christophersen and Gerstenberger, in prep). Five models, EEPAS-0F, EEPAS-0R, EEPAS-1F, EEPAS-1R and PPE (Rhoades and Evison, 2004) were tested in the three-month category using data from 1996 to 2007, and all but EEPAS-1R were shown to be consistent with the data. In a relative comparison test of the five models, the EEPAS-0F model was shown to provide a statistically significant improvement over the other four models. Only one model, M8 (Harte, et al, 2007) was tested in the six-month category and it was shown to be inconsistent with the earthquake data between 1996 to 2007. The five-year models were tested using observed earthquake data from 1984 to 2009 and all three models, a uniform Poisson model, a smoothed seismicity model (Rhoades and Evison, 2004), and the National Seismic Hazard Model (Stirling, et al, 2002), were shown to be consistent with the data. In a comparison test the smoothed seismicity model was able to significantly reject the other two models.

Before testing the M8 model, the model parameters were optimised for the requirements of the testing centre and the code was adapted to conform to the regulations of the NZTC. Finally, a layer of code was written to implement the model within the NZTC environment.

Additionally we report the results of the development of two alternative testing routines for the NZTC. First, we have developed alternatives to the L and N-Tests in CSEP that remove unnecessary calculations and speed up the computational time significantly. Secondly we have developed a protocol for testing long-term forecasts (100 years or more) based on historical Modified Mercalli Intensity information and recorded ground-shaking amplitude data.

1.0 INTRODUCTION

The New Zealand Earthquake Forecast Testing Centre (NZTC) has been developed and implemented at GNS Science (Gerstenberger, 2009), in cooperation with the Collaboratory for the Study of Earthquake Predictability (CSEP) based at the Southern California Earthquake Center. The NZTC is a computational environment and a rule-set for how earthquake forecasts should be tested in New Zealand. The computational environment is largely based on work done by programmers at CSEP with appropriate modifications for the New Zealand setting and New Zealand models. GNS Science has purchased two high-performance servers for operation of the NZTC: a development server that is used for setting up models and ensuring the operation is to the modeller's expectations, and an operations server for retrospective testing and automated prospective testing.

This report is focussed on three aspects of the work of the NZTC: 1) Adaptation and implementation of the M8 model into the testing centre environment; 2) Retrospective testing of New Zealand forecast models in four model classes: one-day models, three-month models, six-month models and five-year models; and 3) Development of alternative testing methodologies including more efficient adaptations of existing testing routines, and a protocol for tests based on forecast ground motion.

One of the difficult problems in earthquake forecasting and seismic hazard analysis is determining how well a particular model performs when compared to observed earthquake data. The difficulty arises because of the small number of large earthquakes that are usually available to compare to the forecasts generated by a model. Not only will this generally limit the power of any tests done but also it means that creative and complicated methodologies are usually required to produce meaningful test results. The most rigorous and unbiased way to test any forecast is through prospective testing, in which the forecast is completely specified prior to the beginning of the test period and the evaluation is completed at the end of the test-period. A disadvantage of prospective testing is that can take a long time to complete, and if the result of testing is that the model contained an obvious flaw which, if known to the modeller at the outset, could have been easily rectified, the time used for testing is mostly wasted.

While prospective tests provide the most robust information about the performance of a model, there is value in supplementing this information with retrospective testing in which a model is tested against data that occurred in the past. Retrospective testing does not avoid the inevitable bias that arises when a model is tested on the same data from which it was developed, but it can reveal when a model is performing very poorly and situations (e.g., regions) in which a model may perform better. If the reason for poor performance is merely a technical error in the model's implementation, the modeller then has the opportunity to correct it before the prospective testing is too far advanced.

Conducting retrospective tests according to the prospective–testing protocols can also give insights into practical inefficiencies in the NZTC software and limitations and deficiencies in the presentation of the core test results in the present system. These insights can be used to identify possible alternative testing methodologies and to improve both the efficiency of the software and the clarity of the test results when the system is upgraded in the future.

2.0 M8 MODEL IMPLEMENTATION

2.1 Original Model

The original M8 algorithm (Kellis-Borok & Kossobokov, 1990) is intended to predict large magnitude events (typically magnitude 8 -- hence its name) within a specified region, the size of the region being determined by the target magnitude. Briefly, the (declustered) catalogue for the region is examined at (typically) 6-month intervals for a period of at least 10 years in order to set threshhold values for various characteristics of the seismicity of the region. The threshold values are used in later 6-month intervals to identify Times of Increased Probability (TIPs) from the empirical distributions generated.

2.2 SSLib Implementation

The model is written in the R Programming language and utilises the Statistical Seismology Library (SSLib) package (Brownrigg & Harte, 2005) The original SSLib implementation involved two modifications to the algorithm. The first was to introduce a target magnitude which was different from the magnitude used to determine the size of the region of interest. The second was to use a grid of overlapping circular regions to cover a much wider area. In this implementation, each of the overlapping areas was considered independently. A subsequent modification of the SSLib algorithm introduced the concept of synoptic forecasts, whereby the information from the overlapping circles is combined statistically in order to produce probability forecasts for non-overlapping cells.

2.3 Modifications for CSEP

The M8 algorithm as developed within SSLib was primarily an investigative tool, rather than a predictive tool. A number of modifications were required in order to fit it into the CSEP framework. Firstly, the algorithm assumes that everything is based on 6-month intervals and so the time points used in the calculations were assumed to be on 6-month calendar boundaries. Further, the declaration of TIPs was assumed to apply to the 6-month period following the last boundary. For the CSEP testing framework, the software must make no assumptions about the end date of the catalogue provided, nor about the length of time for which a prediction is to be made (although this is typically 3 or 6 months for this type of algorithm). Finally, the output from the synoptic forecast of the SSLib software, which is a probability forecast for events of a specified magnitude or greater, needed to be converted into individual probabilities for each magnitude bin from 5.0 to 9.0 in intervals of 0.1.

The final stage of implementation involved creating a wrapper around the M8 code in the python programming language and ensuring that the model interacted appropriately with the CSEP environment.

3.0 RETROSPECTIVE TESTING

All retrospective testing, except where detailed below, was performed within the CSEP testing environment. This environment consists of highly specified procedures that allow for optimal reproducibility of any testing procedure and archiving of all data and computer code used in the tests. All control of the model is given to the NZTC and no interaction from the modeller is allowed. Additionally, all testing and data acquisition is fully automated which reduces the chance for human error in the procedure. For full detail of how the testing centre

is specified, see Schorlemmer and Gerstenberger (2007). While testing within the CSEP environment allows for the most rigorous and transparent results, it comes with a cost of a large computational overhead. For prospective testing, the computational overhead is minimal when compared to the time-periods of testing and causes no problems; but for retrospective testing, where many testing periods are evaluated back-to-back, the overhead greatly slows the testing procedure.

3.1 The Tests

We have used the core tests as implemented within CSEP (Schorlemmer, et al, 2007). These tests are developed to evaluate earthquake rates, not ground motion, and test magnitudes in 0.1 M units over a range of magnitudes such as 5.1, 5.2, 5.3, up to 9.0. Spatially the tests cover all of New Zealand, including an offshore buffer (Figure 1). The test region is subdivided into 0.1 degree square cells for evaluation. The time periods for evaluation vary for each model class and are discussed in the *Model Classes* section.

The three statistical tests aim to evaluate the consistency of the forecast with the observed data and are the N-Test, L-Test and R-Test. The N-Test is the most basic of the three and includes no spatial information. This test compares the total number of observed events in the entire time period and region to the total number forecast; it computes the p-value of the observed number of earthquakes within the distribution for the number of earthquakes given from the model. The L-test extends the N-Test to include information on earthquake magnitude and location contained in the expected number of events in each cell under the model, and again compares the actual outcome with that predicted by the model, using the likelihood statistic; it computes the p-value of the actual likelihood statistic within the distribution of possible likelihoods given the model. Through examining the results of the N-Test and L-Test together, one can gain an understanding of the consistency of the forecast cell expectations with the actual observed data. Finally, the R-Test is performed. It evaluates the ratio of the likelihoods of the two models, and compares it with the distribution of this ratio given each model in turn. For a complete mathematical description of the testing procedures please see Schorlemmer et al (2007).

No test on its own is able to give a complete understanding of the performance of the models; the tests are designed to test different aspects of model performance. It is fully possible for a model to pass one test and fail another. By evaluating the three tests together we aim to give a more complete understanding of the performance of the models.

3.2 Model Classes

Not all forecasting models target the same information. Some models aim to forecast the location and magnitudes of all aftershocks following a large main shock, other models aim to forecast large earthquakes over a time period that might be useful in developing building codes. To best learn about the performance of the models, separate testing procedures are needed for models with different goals; we term these procedures Model Classes. Within the NZTC we have 4 model classes: one-day models, three-month models, six-month models, and five-year models. Table 1 shows the four model classes, the models within each class and the relevant time-periods.

3.2.1 One-day-models

Typically one-day models aim to forecast the locations and magnitudes of aftershocks and put less emphasis on forecasting main shocks. These models estimate earthquake occurrence for magnitudes $4.0 \leq M \leq 9.0$ for each 24-hour time-period. In the retrospective testing it was intended to test these models every day for a 10 year time-period; however, doing so would have taken approximately 6 months of computational time within the CSEP environment. Therefore, the two primary models, STEP (Gerstenberger, et al, 2005) and ETAS (Rhoades et al, 2008) were evaluated within the CSEP environment for approximately 6 months of historical data. STEP, the Abundance model as implemented within STEP (Christophersen, 2005), and a reformulation of STEP to improve basic assumptions in the model (Christophersen and Gerstenberger, in prep) were evaluated outside of the CSEP environment, but using identical testing routines, for the period 1996 to 2006. In the proposal it was specified that the STEP+EEPAS model would be implemented in the testing centre; this has proven to be more difficult than expected and is not yet implemented, but will be included in a future iteration of the testing centre system.

3.2.2 Three-month-models

The three-month models are tested for magnitudes $5.0 \leq M \leq 9.0$. Every three months the model generates a new forecast that is allowed to use all data that occurred up to the time of the start of the forecast. This model class was evaluated every three months from 1996 to the end of 2006. Five models have been implemented in this model class; four of these are variations of the EEPAS model as described in Rhoades, et. al., (2009) and the fifth is a smoothed seismicity model, PPE (Rhoades & Evison, 2004).

3.2.3 Six-month-models

Initially the NZTC did not contain any six-month-model class and the other CSEP-based testing centres do not intend to implement a six-month class. This class is implemented to allow for fair testing of the M8 model (Harte, et al, 2005). It is not intended to create a new model class for every new model that is presented to us; however, after discussion with the modellers it was clear that the M8 model would be significantly disadvantaged if forced into either the three-month or the five-year class and because of the model's historical significance it was desirable to include a class that would fairly represent the model. As with the three-month class, the six-month class was tested for magnitudes $5.0 \leq M \leq 9.0$, and the model was evaluated every six months for the time period of 1996 to the end of 2006. The M8 model is the only model in this class at present, but it is planned to implement additional simple models (e.g., smoothed seismicity and uniform Poissonian) in order to gain an understanding of the M8 model's relative performance.

3.2.4 Five-year-models

This class is primarily targeted at long-term models such as the New Zealand National Seismic Hazard Model (NSHM; Stirling, et al, 2002). Because testing such a model over its intended life-span (e.g., 50 to thousands of years) is infeasible, we have chosen to scale the models down to five years and test them for this shorter time period. This is primarily an issue for prospective testing (i.e., a scientific procedure that will take 50 years to finish is beyond the career of most scientists and the results will likely be irrelevant upon completion); for retrospective testing we can string together multiple five-year periods to get a better reflection of the model's performance over its intended time period. Because this class

requires relatively few evaluations, computational time is not a limiting factor when determining the testing period. We chose to test every five years from 1984 to the end of 2008. Extending the test to earlier dates would have increased the magnitude of completeness in some areas to a value greater than than required by the tests and would have limited the usefulness of the results; completeness is required to be under five for a the regions shown in Figure 1. The five-year class was tested for $5.0 \leq M \leq 9.0$. Four models have been implemented in this testing class: the NSHM, a uniform Poisson model (i.e., the rates are the same for each grid node and based on the total expected rate from the earthquake catalogue up to 2005), the PPE model (Rhoades & Evison, 2004) which is a smoothed seismicity model based on the earthquake catalogue up to 2005, and a PPE model based on the synthetic seismicity catalogue for the Wellington region from the model of Robinson and Benites (1996). However, due to an error in processing, the results for the latter model are erroneous and will need to be recalculated at a later date. For this reason they are not presented here.

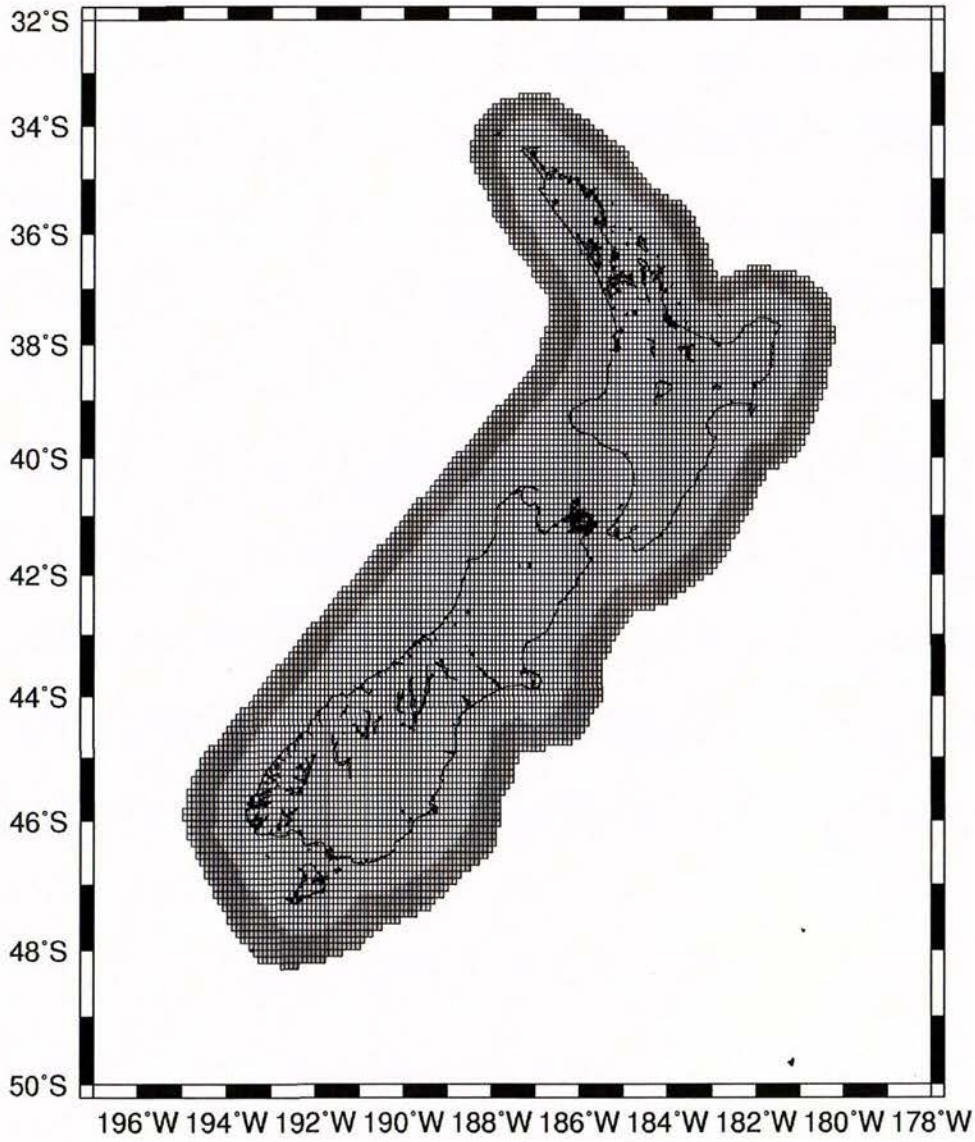

Figure 1 Models are evaluated using events occurring in the testing region which is shown in pink. Models are supplied events from the slightly larger area, the collection region, which includes both the grey and the pink regions.

The NSHM contains two types of sources: grid-based rates and fault-based rates. Testing fault-based rates directly is a problem that has not yet been solved; all tests in the NZTC are grid-based. For this reason we have distributed all fault-based rates in the NSHM over 0.1 degree square grid cells for the purpose of testing the NSHM in the NZTC. While this retains the same total number of forecast events for the model, it is not strictly the NSHM and we will refer to it as the NSHM-GB. A particular problem with the interpretation has to do with the recurrence interval of large events in the NSHM. In our interpretation, even though the correct total number of events is retained for each single fault, the rupture rates at each grid cell are underestimated, when compared to what is expected for the fault, by the number of grid cells per fault. For example a fault with a recurrence interval of 100 years that passes through 10 grid cells will result in a 1000 year recurrence interval for each of the 10 grid cells.

Table 1 Models classes, the time-periods evaluated for each class, and the models within the class.

	1984-2009	**1996-2006**	**1996**
One-day-models		STEP, Abundance, New STEP (no R-test)	STEP & ETAS
Three-month-models		EEPAS-0F, EEPAS-0R, EEPAS-1F, EEPAS-1R, PPE	
Six-month-models		M8	
Five-year-models	NSHM-GB, SUP, PPE		

4.0 THE DATA

All tests discussed in this report were performed using data automatically downloaded from the GeoNet database. For the five-year-model class the data was declustered using the default parameters of the Reasenberg (1985) declustering algorithm; this is the same method used in developing the NSHM-GB. For the other model classes, all data downloaded from the catalogue was available for use in the testing.

5.0 RETROSPECTIVE TESTING RESULTS

5.1 Five year models

In this class the model codes are not required to be installed in the NZTC and only a static forecast for any five-year period is supplied. The supplied forecasts for the NSHM-GB and PPE are shown in Figures 2 and 3.

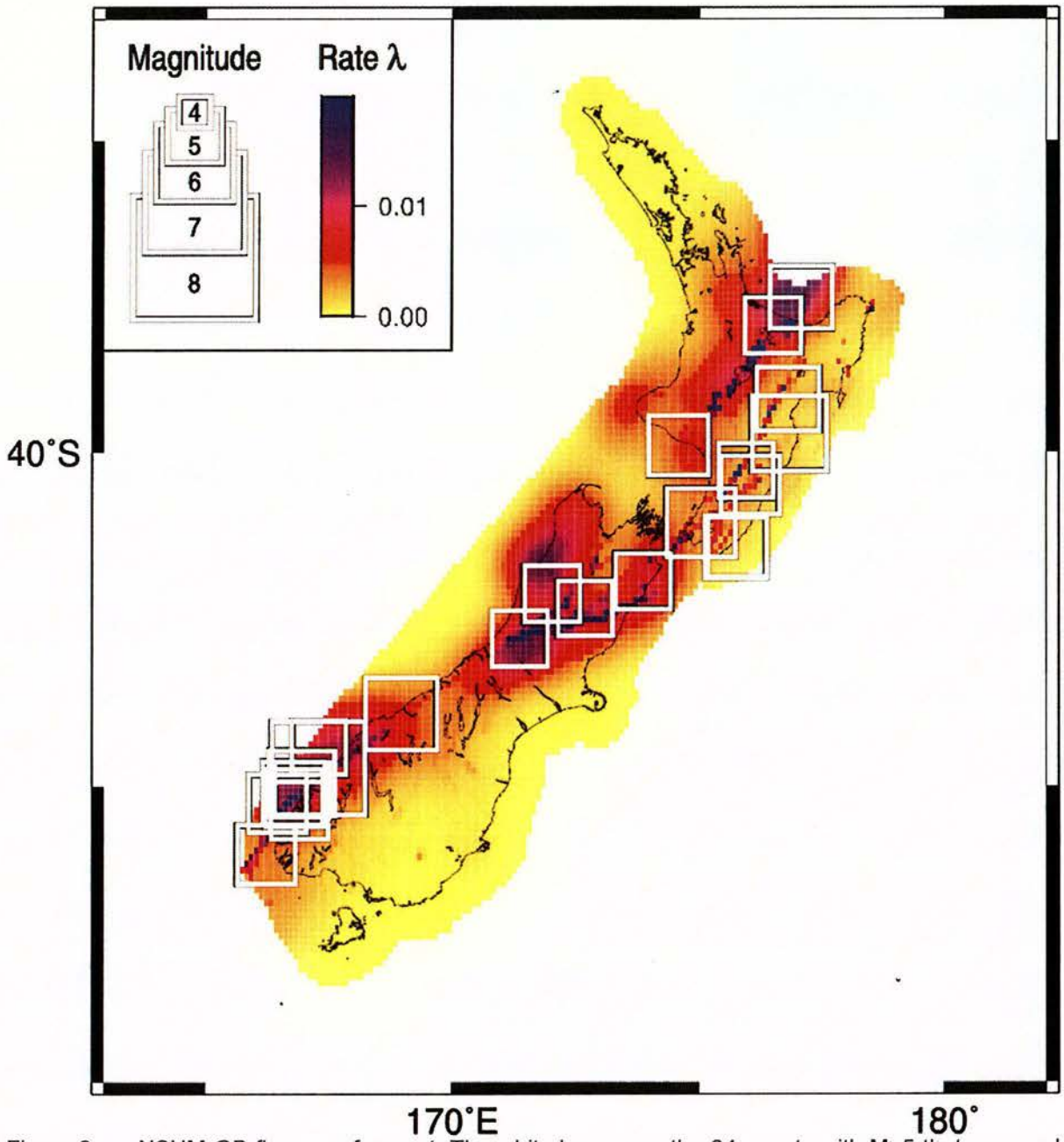

Figure 2 NSHM-GB five-year forecast. The white boxes are the 24 events with M>5 that occurred between 2004 and 2009.

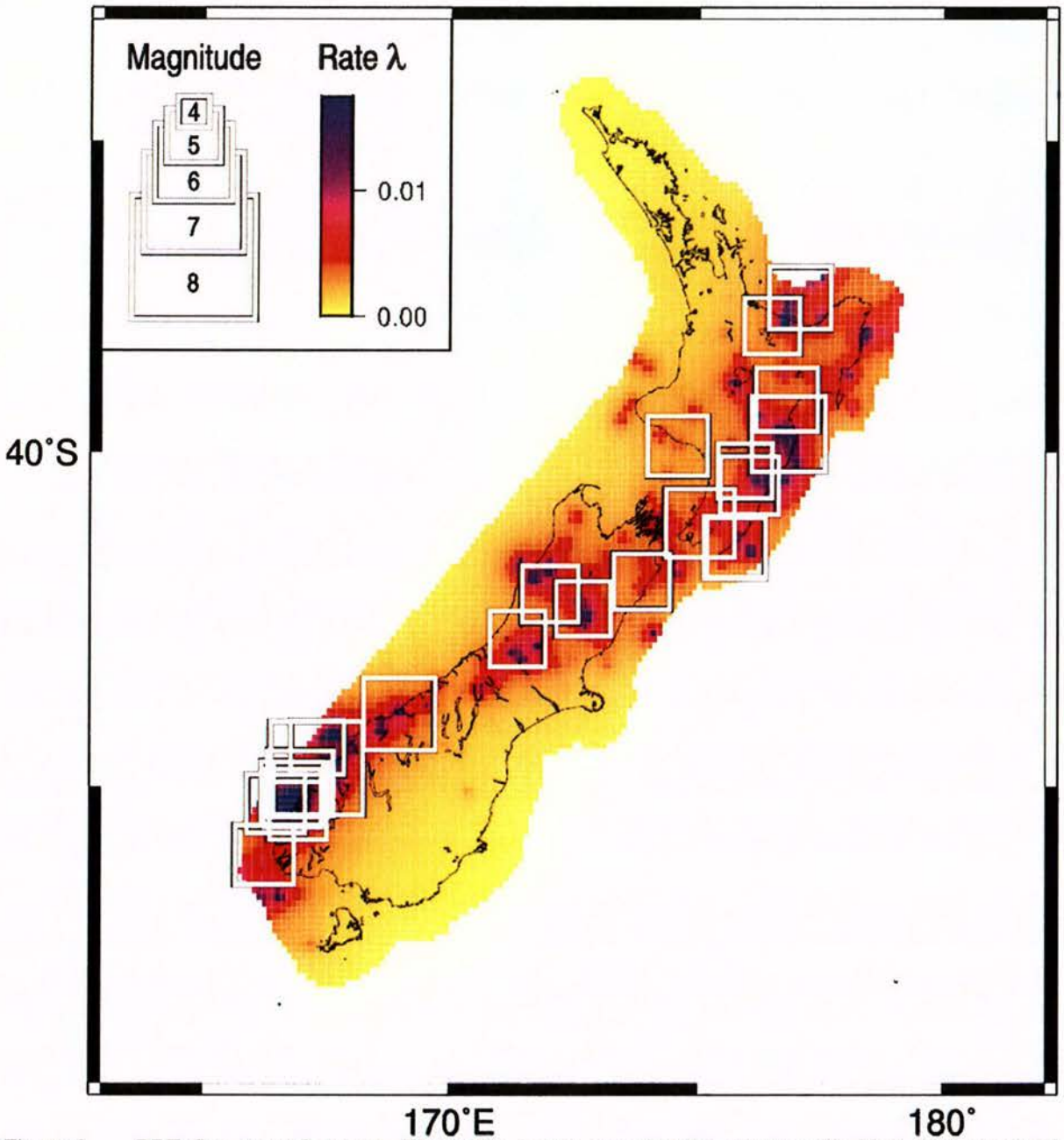

Figure 3 PPE five year forecast. The white boxes are the 24 events with M>5 that occurred between 2004 and 2009.

5.1.1 Five-Year N-Test Results

The N-Test is a basic test of the total number of events forecast by the model, ignoring all spatial information. The result of the test from a single five-year period for the NSHM-GB is shown in Figure 4. The green curve shows the cumulative distribution of numbers of events that can be considered consistent with the forecast assuming the model is Poissonian. The solid vertical black line shows the actual number of observed events for the time period. If the observed number falls somewhere in the middle of the green curve the model cannot be rejected as being inconsistent with the observation. To quantify this, two grey bars are added to the graph at probabilities of less than 2.5% and greater than 97.5%; if the green curve contacts either one of these grey areas the model can be rejected as being inconsistent with the observed data (i.e., it is unlikely to be observed given the forecast). In Figure 4, the forecast is consistent with the data.

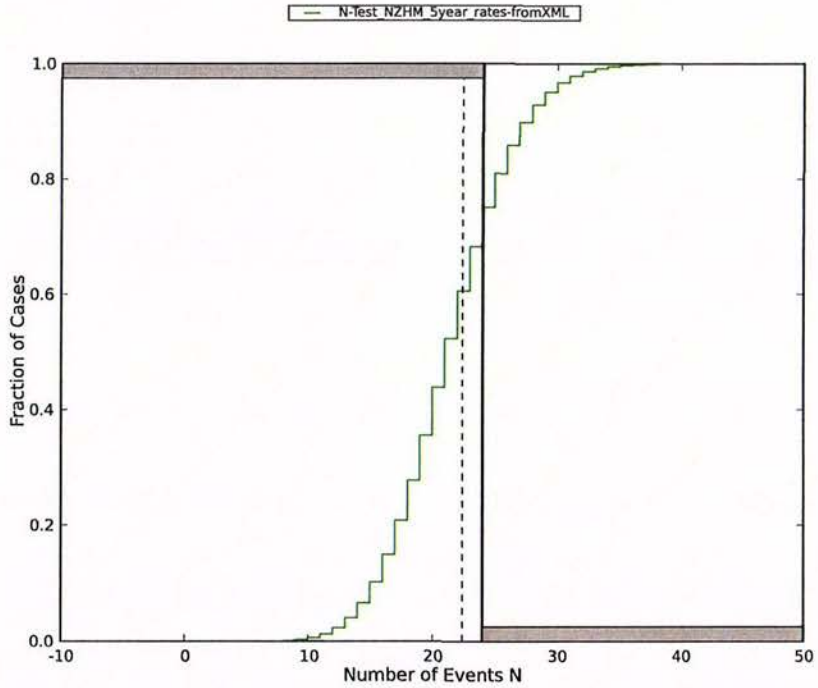

Figure 4 Five Year N-Test of the NSHM-GB 2004-2009

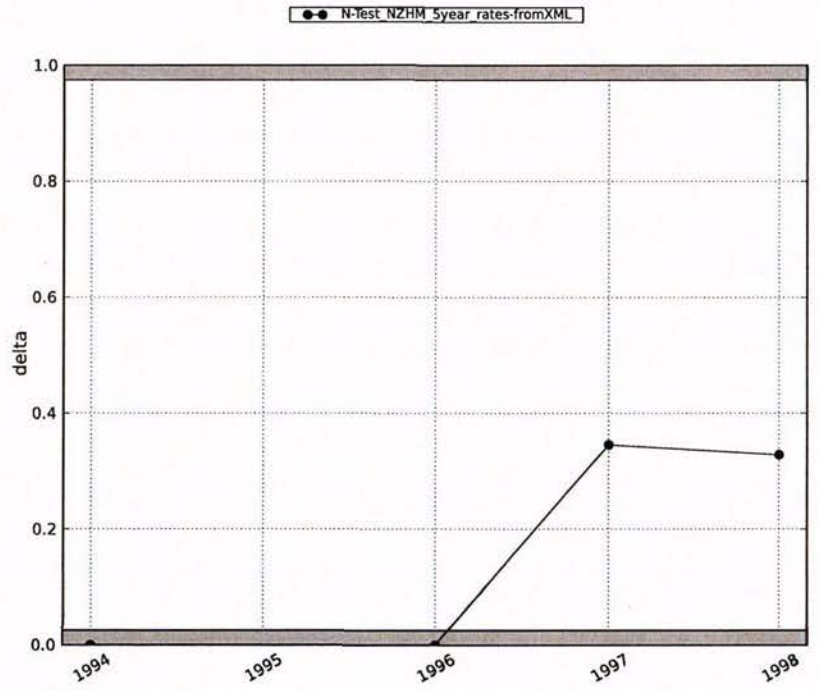

Figure 5 Intermediate N-Test for the NSHM-GB for 1994-1999

Figure 5 shows how the N-Test results evolved over a single five-year period for the NSHM-GB. On the Y-axis, Delta, reflects the single point, from a plot such as shown in Figure 4, where the observed value (i.e., the black vertical line in Figure 3) crosses the green curve (which displays those numbers consistent with the forecast). From the single point we can understand if the forecast is consistent with the model or not; if Delta is between .025 and .975 the model may not be rejected as being inconsistent with the observations. If Delta is greater than .975 the model is under-predicting the observed number of earthquakes; if Delta

is less than .025 the model is over-predicting the observed number. In Figure 5, the forecast rates are scaled to the appropriate time-period (e.g. 1 year, 2 years, etc) and the forecast appears to initially be over-predicting the number of events; however, for the complete five year period, the forecast is consistent with the data with 22 observed events and 22.39 forecast.

Figure 6 shows the cumulative 25 year N-Test result for the NSHM-GB (1984-2009). As shown in Table 2, for the total time period the NSHM-GB model forecast 113 events and 112 were observed. All three models (NSHM-GB, SUP and PPE) in this class were remarkably similar in total number of events forecast for the 25 year time period and none of the forecasts could be rejected. Table 2 shows the number of forecast events for each five-year time-period for each model and compares them to the observed numbers of events.

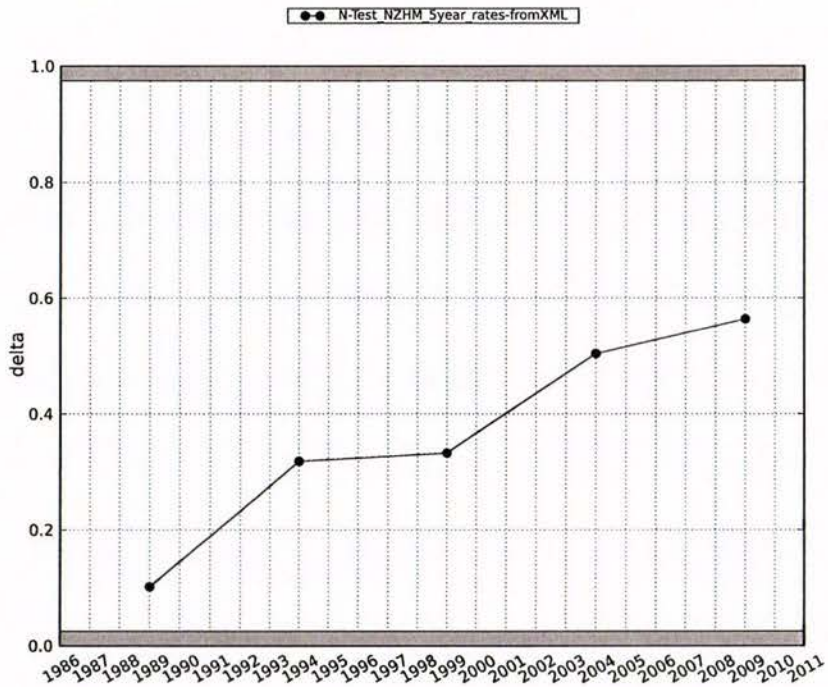

Figure 6 Cumulative N-Test for NSHM-GB 1984-2009

5.1.2 Five-year L-Test Results

The next test performed is the L-Test which, beside total numbers of events, includes spatial information (i.e., the location of the forecast events). Figure 7 shows the results of the L-Test for 1994-1999 for the NSHM-GB. Similar to the Delta of the N-Test, the gamma of the L-test represents a single point taken from each individual test (e.g., for a 1-year time-period within a 5-year test). In the case of the L-Test, gamma represents the point where the line representing the likelihood of the observed earthquakes crosses the curve which represents the distribution of likelihoods that are consistent with the forecast. For complete details of the test, see Schorlemmer, et al (2007). Unlike the N-Test, this evaluation takes into account the forecast for every grid-cell within the test region and for each discrete magnitude within the tested range. Like the N-Test, the L-Test is a two-sided test, meaning that a model can be rejected if the observed likelihood is "too high" (gamma < .025) or "too low" (> .975). However the likelihood can be "too high" in a number of special circumstances

(Schorlemmer, et al, 2007) and a model should not be rejected on this result alone; the results should be confirmed with the N-Test. Like the N-Test, the L-test shows the NSHM-GB to be initially inconsistent with the data, but, finally consistent with the observed data once the entire time-period is complete.

The complete L-Test for the 1984-2009 time period is shown in Figure 8. As with the N-Test the NSHM-GB is shown to be consistent with the observed data and cannot be rejected. As with the N-Test also, neither of the other two models can be rejected in the L-Test so are not shown to be inconsistent with the data.

Table 2 Number of forecast events for the three five-year models for each of the 5 testing periods. The forecast misfit (Observed-Forecast) is shown. Cells shown in grey are over-predictions; all other cells are under-predictions. No five year model is rejected in the intermediate or cumulative N-tests.

Model & 5 year forecast	**NSHM – 22.4**	**PPE – 20.6**	**SUP – 21.5**
1984-1989: 16	-6.4	-4.6	-5.5
1989-1994: 25	2.6	4.4	3.5
1994-1999: 22	-0.4	1.4	0.5
1999-2004: 26	3.6	5.4	4.5
2004-2009: 24	1.6	3.4	2.5
Total misfit	1	10	5.5

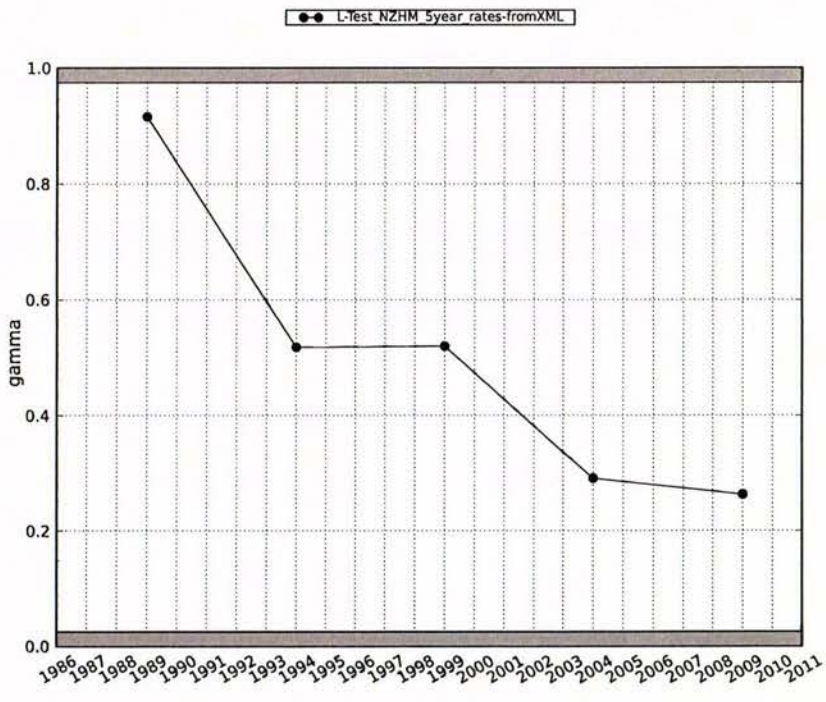

Figure 7 Intermediate L-Test of the NSHM-GB for 1994-1999

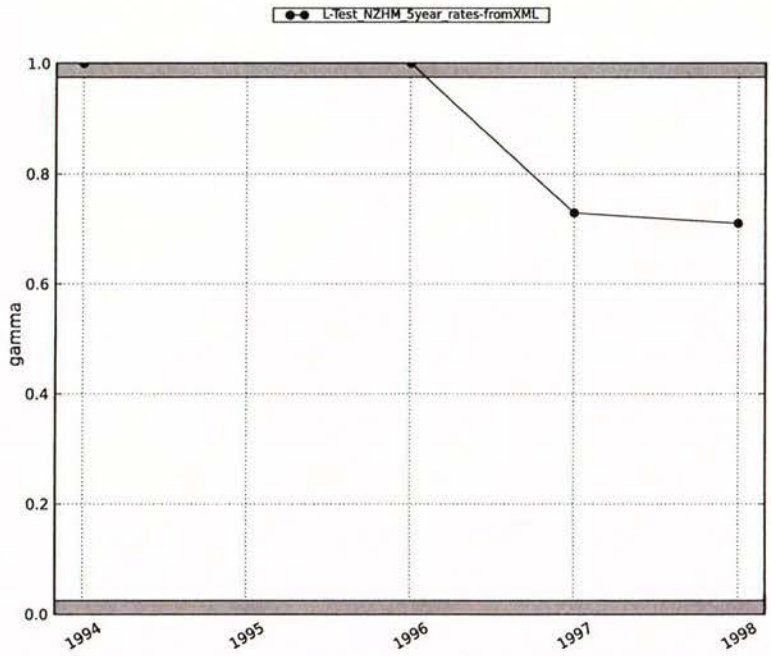

Figure 8 Cumulative L-Test for the NSHM-GB 1984-2009

5.1.3 Five-year R-Test Results

The final test of the five year models is the R-Test which compares each model to all others. In the R-Test we aim for a relative comparison of two models to gain an understanding of whether one model has performed better than the other. Figure 9 shows the results of the R-Test comparing the NSHM-GB to the SUP model. Two results are shown in the plot: one assuming the NSHM-GB is the null hypothesis and another assuming the SUP model is the null hypothesis. For full details of how the R-Test is calculated please see Schorlemmer et al (2007). Like the other tests, the R-test rejects the model as being inconsistent with the observed data if the result is within the rejection bars (>.975 or <.025). Figure 9 shows that both models are rejected based on the other model being used as the null hypothesis.

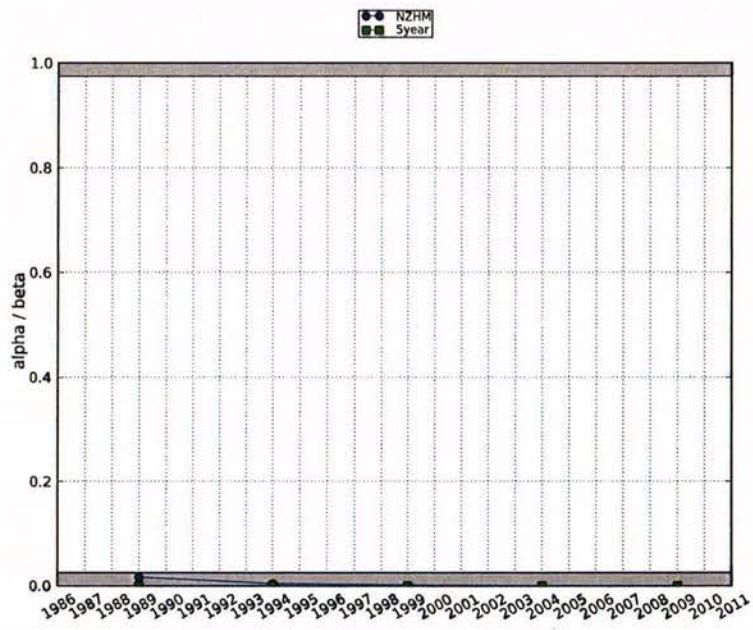

Figure 9 Cumulative R-Test of NSHM and SUP, 1984-2009. Note that both sets of results plot in the bottom rejection bar.

Figure 10 shows the R-Test comparing the NSHM-GB to the PPE model. In this test, for each of the 5 year time periods, the NSHM-GB can be rejected when compared to the PPE model. In the comparison of the PPE model and the SUP model, both models can be rejected. While possibly seeming contradictory, these results indicate that each model contains some information that the other does not, and this gives the model an advantage in forecasting some earthquakes. On its own, each model is shown to be consistent with the data in both the L-Tests and the N-Tests. In the R-Test the PPE model is shown to be probably the best forecast model for the time period tested.

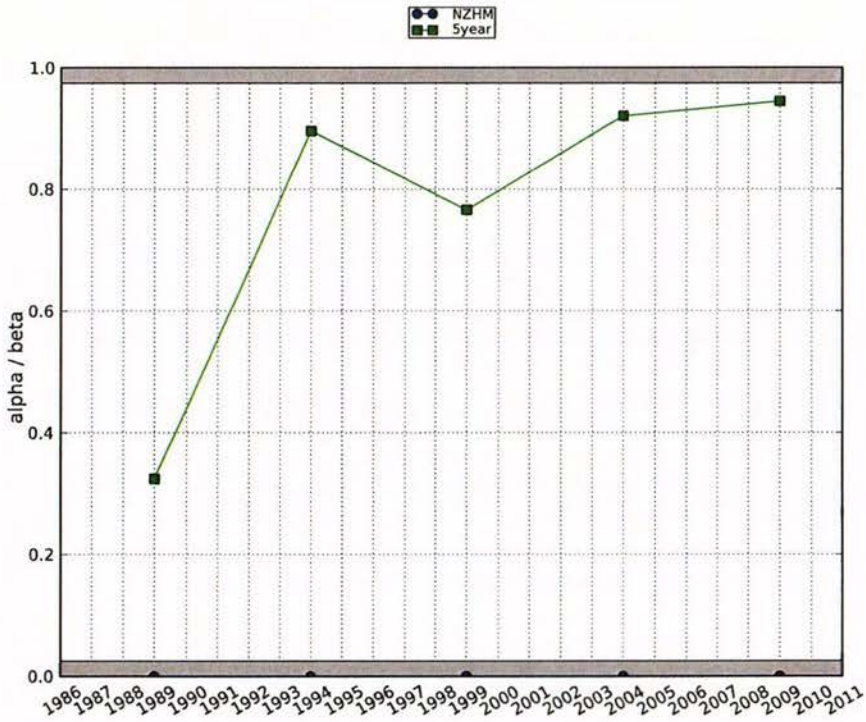

Figure 10 Cumulative R-Test of NSHM and PPE, 1984-2009

The advantage of the PPE model over the NSHM-GB probably lies in the spatial location of the forecast earthquakes, however this gain is not significant when compared to the spatially uniform SUP model. The significant advantage over the NSHM-GB model may come from the fact that the NSHM-GB model focuses the largest earthquake probabilities on mapped faults with a smaller amount going off-fault. In the time-period covered in this test, most of the observed earthquakes did not occur in cells intersected by mapped faults.

5.2 Six-month Models

Testing of the six-month class is limited to the M8 model so only the N-Test and L-Test have been performed. For prospective-testing it is planned to include appropriate PPE and SUP models so that R-tests comparison can also be made. A typical 6-month forecast for the M8 model is show in Figure 11; one event occurred in this time period and is shown in the white box.

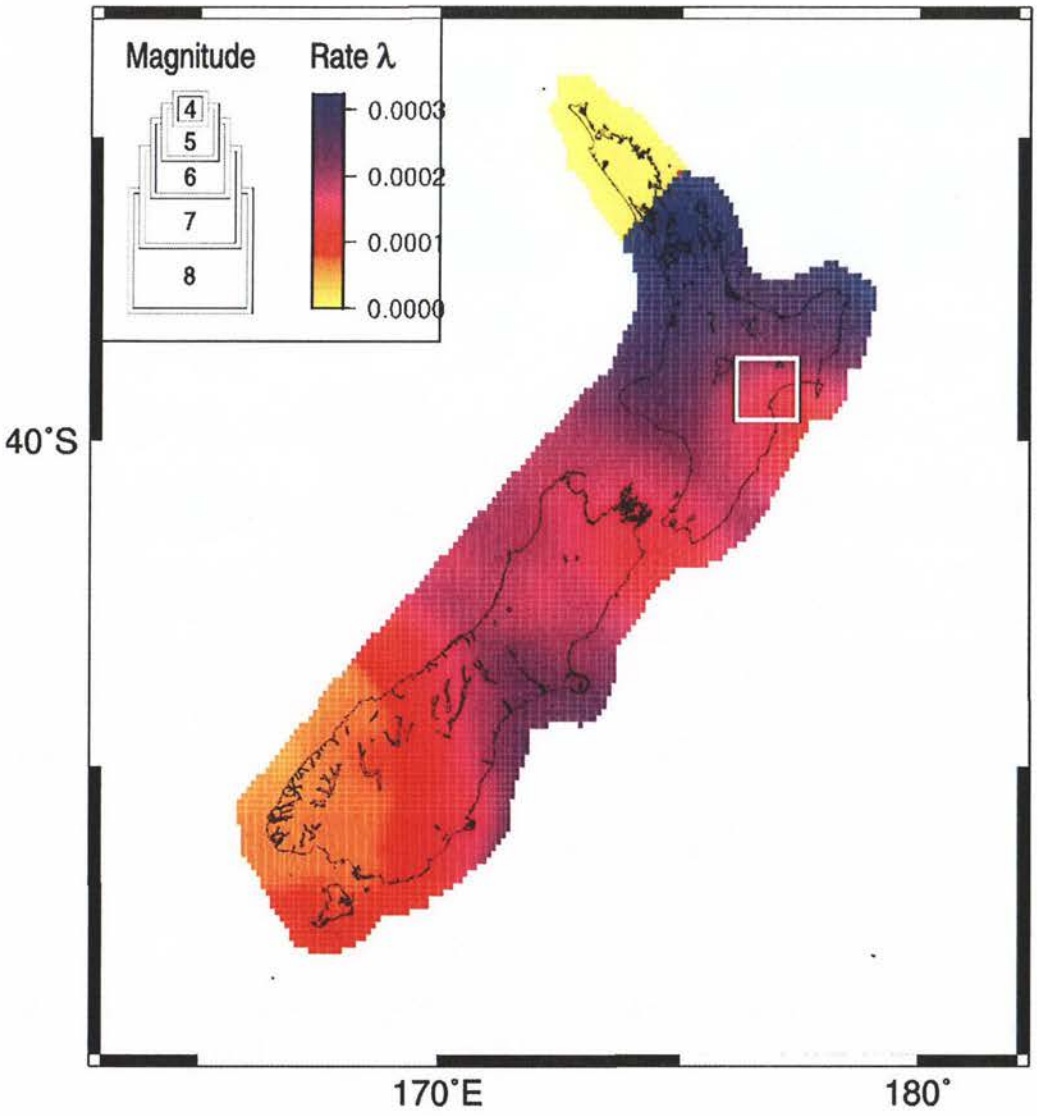

Figure 11 M8 Forecast 1-7-2006 through 31-12-2006. Observed events M>5 marked by white box

5.2.1 Six-month N-Test Results

Figure 12 shows the cumulative N-Test results for 1996-2007. For the time period from 1996 to 2003, the forecasts were consistent with the data; from 2003 to 2007 the model under-predicts the observed number of events. A breakdown by 6-month time period is shown in Table 3. From this table it is clear that a failure to predict the large aftershocks to the 2003 and 2005 Fiordland events is what caused the failure of the model.

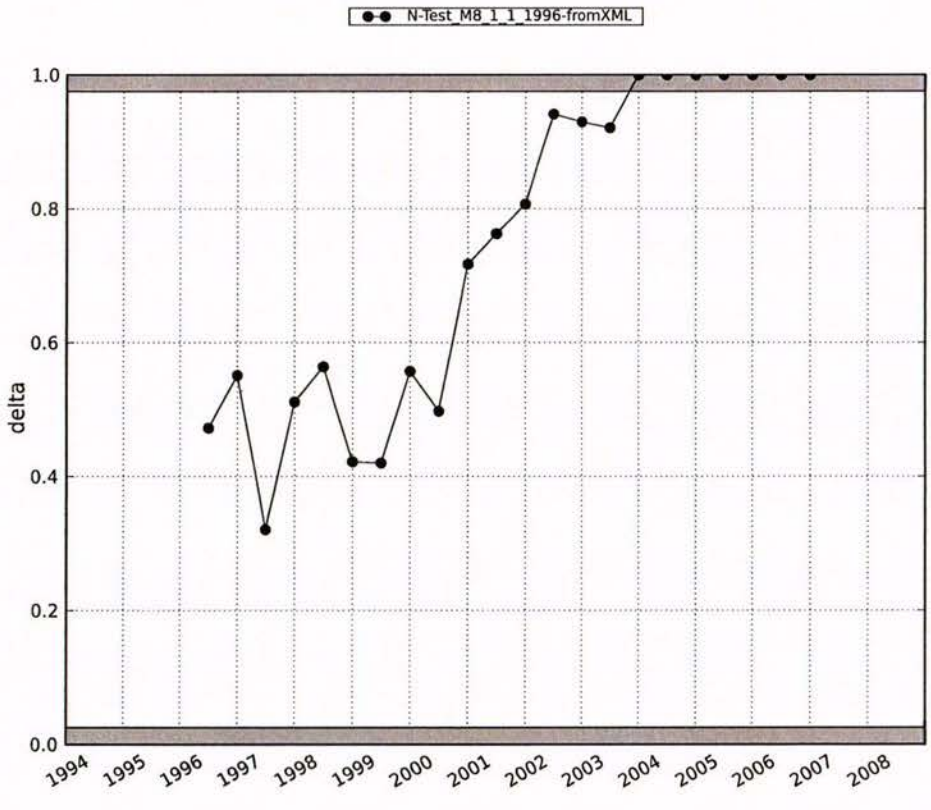

Figure 12 Cumulative N-Test results for the M8 model for 1996-2007.

Table 3 M8 forecast and observed for 1996-2006. Delta is the p-value of the observed number of earthquakes during the time-period within the distribution of the number earthquakes under the M8 model; a forecast with .025 < Delta < .975 cannot be rejected as inconsistent with the observations.

Time period	Observed	Forecast	Delta
1996; Jan 1 to June 30	1	1.77	0.47
1996; July 1 to Dec 31	2	1.67	0.77
1997; Jan 1 to June 30	2	3.41	0.34
1997; July 1 to Dec 31	4	2.73	0.86
1998; Jan 1 to June 30	3	2.53	0.75
1998; July 1 to Dec 31	1	2.3	0.33
1999; Jan 1 to June 30	2	2.07	0.66
1999; July 1 to Dec 31	3	1.58	0.92
2000; Jan 1 to June 30	1	1.64	0.51
2000; July 1 to Dec 31	4	1.29	0.99
2001; Jan 1 to June 30	2	1.23	0.87
2001; July 1 to Dec 31	2	1.14	0.89
2002; Jan 1 to June 30	5	1.17	1.00
2002; July 1 to Dec 31	1	1.13	0.62
2003; Jan 1 to June 30	1	1.18	0.67
2003; July 1 to Dec 31	27	1.08	1.00
2004; Jan 1 to June 30	2	1.29	0.86
2004; July 1 to Dec 31	3	1.14	0.97
2005; Jan 1 to June 30	10	1.21	1.00
2005; July 1 to Dec 31	1	1.08	0.71
2006; Jan 1 to June 30	0	1.08	0.34
2006; July 1 to Dec 31	1	1.00	0.74
TOTAL	78 (52 excluding Fiordland aftershocks)	34.72	

5.2.2 Six-month L-Test Results

The L-Test shown in Figure 13 shows similar results to the N-Test with the forecasts of the M8 model shown to be consistent with the observed data up until the occurrence of the 2003 Fiordland event. From this time to the end of the test, the cumulative performance of the model is shown to be inconsistent with the observed data.

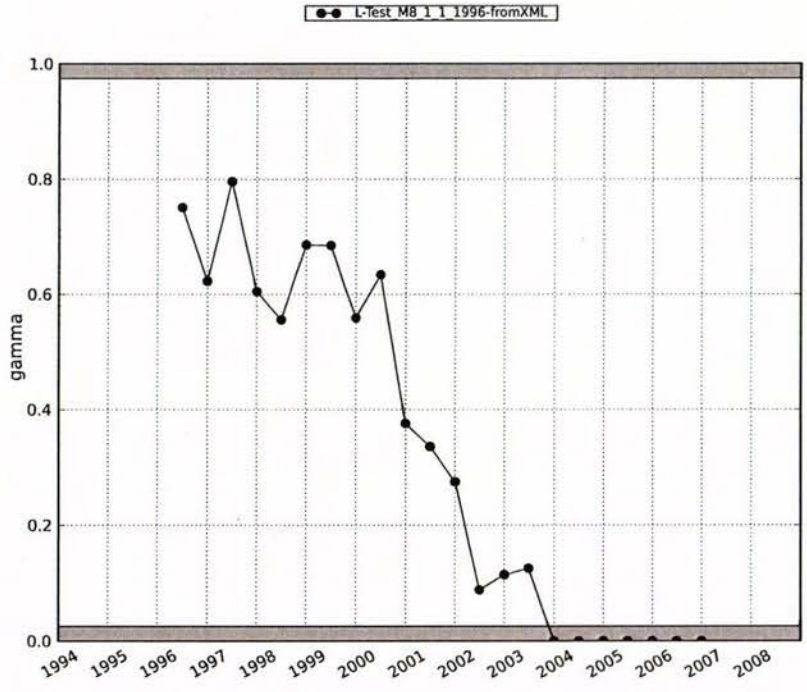

Figure 13 Cumulative L-Test for M8 for 1996 through 2007

5.3 Three-month Models

The three-month model class was tested from 1996-2007, resulting in 48 3-month testing periods. In this class four variations of the EEPAS model were tested as was a 3-month version of the PPE model. An example forecast of the EEPAS-0F model for 7-1-2006 to 10-1-2006 is shown in Figure 14. One event with magnitude greater than 5 occurred during this time period and its location is shown in the white box on the map.

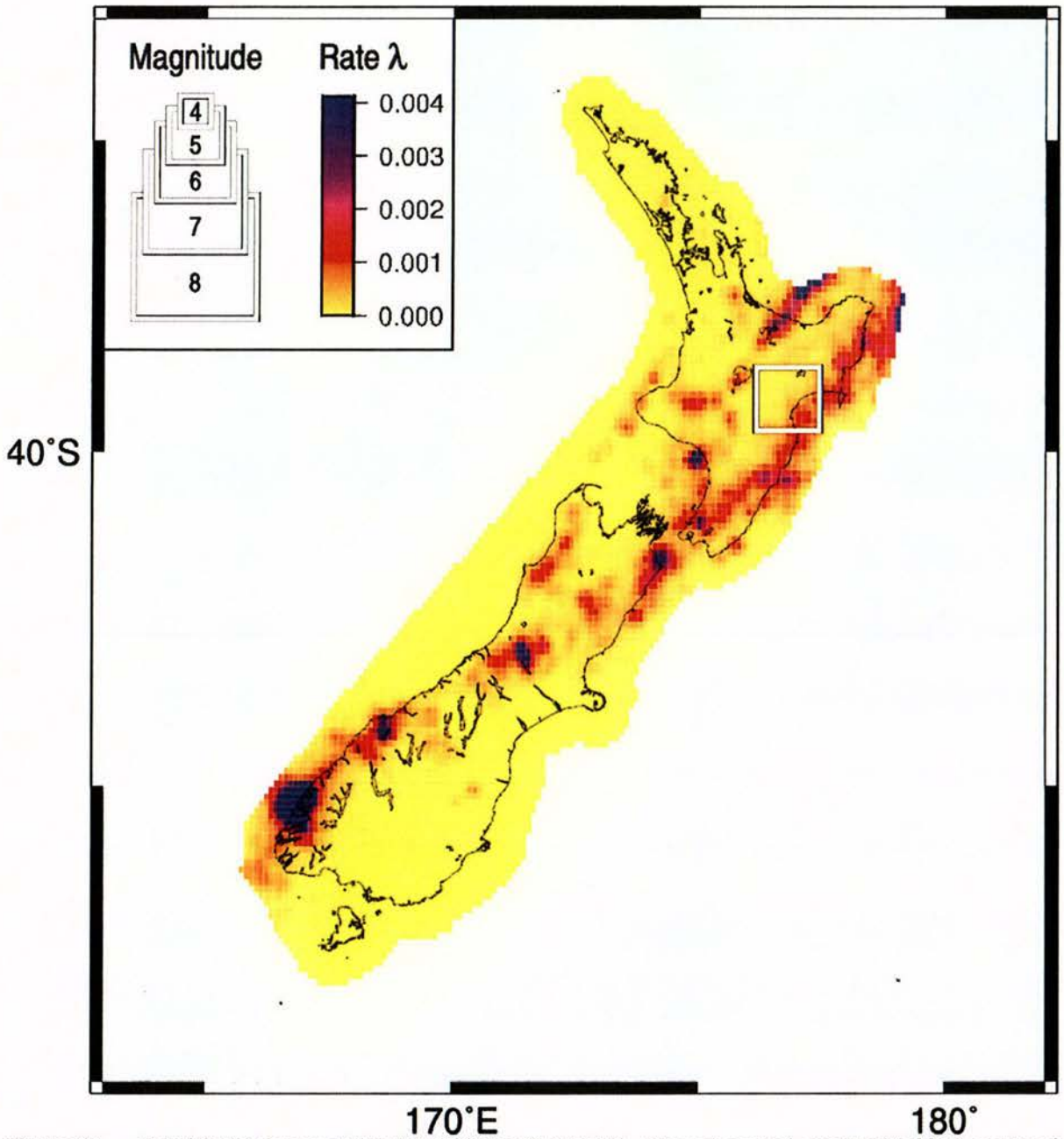

Figure 14 EEPAS-0F Forecast for 7-1-2006 to 10-1-2006. One observed event with M>5 occurred in this time period and is shown in white box.

5.4 Three-month N-Test Results

The cumulative N-Test results for the EEPAS-0F model are shown in Figure 15. It can be seen that the model started out consistent with the observed data, but by 2000 the model appeared to be over-predicting the observed data; however, in 2003, following the occurrence of the Fiordland earthquake to the end of the testing period in 2007, the cumulative results of the N-Test for the EEPAS-0F model indicate that the model is consistent with the data. The cumulative N-Test results for the remaining three EEPAS models and for the PPE model show remarkably similar test results and no model is rejected based on the N-Test.

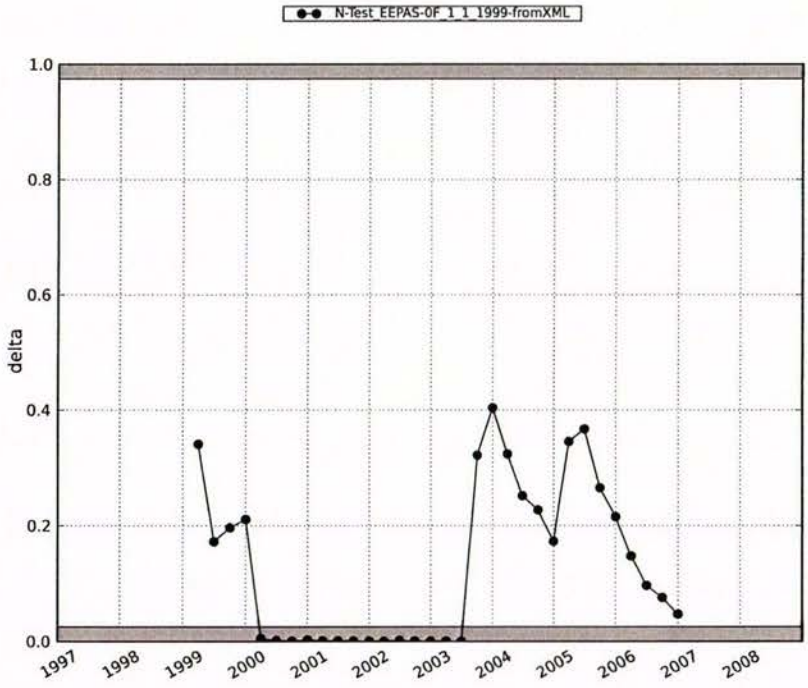

Figure 15 Cumulative N-Test results for the EEPAS-0F model for 1996-2007.

Figure 16 shows a breakdown of the N-Test results for the EEPAS-0F model for each 3-month time period compared to the forecast misfit (Observed-Forecast) on the X-axis; the results shown in this figure are non-cumulative and represent only the performance of the model in a single three-month time period. As with the cumulative plots, Delta represents the point where the line representing the observed number of events intersects the curve representing the numbers of events that can be considered consistent with the forecast. In only three of 48 time-periods tested can the EEPAS-0F model be rejected. In two of these cases the EEPAS-0F model greatly under-predicted the number of expected events. This is due to the fact that aftershocks were not predicted such as for the Fiordland event shown at the far top right of the plot where the EEPAS-0F model under-predicted by 22 events for the 3-month time period. In one case the model was rejected for over-prediction when the forecast contained 10 events more than were observed. Overall this plot shows that the model is consistent with the observations, when considering number of events only, and only in time-periods with an extreme number of events (low or high) can the model be rejected.

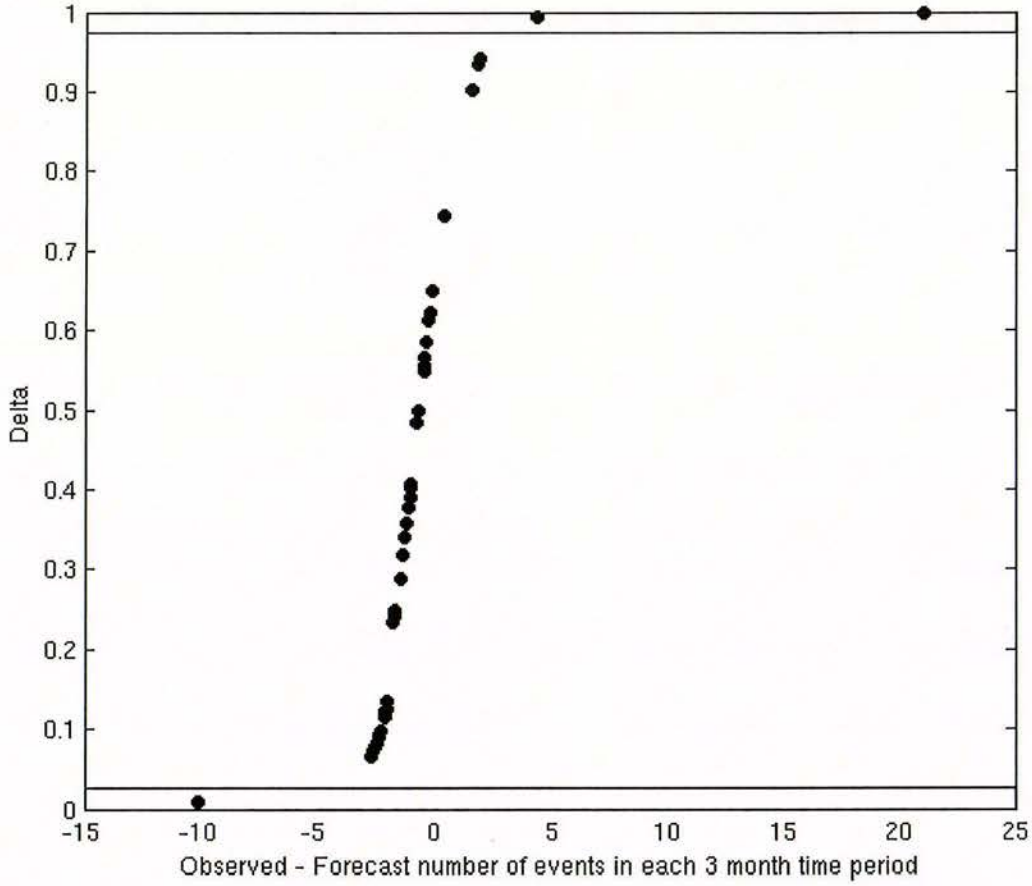

Figure 16 Breakdown of EEPAS-0F N-Test results for each time period. The X-axis shows the forecast misfit: Observed-forecast number of events and the Y-Axis shows Delta, which is the significance of the forecast; those forecasts falling within the grey regions are rejected.

In a similar examination of the PPE model, it can be rejected in only two time-periods in the N-Test, and in both cases it is under-predicting the total number of events.

5.4.1 Three-month L-Test Results

Figures 17 and 18 show the cumulative L-Test results for the EEPAS-OF model and the PPE model for the entire time period. Again, the results are very similar to the N-Test results with the models judged to be consistent with the data for the first few years, but gradually the cumulative performance is indicated as inconsistent with the data by 2000. As with the N-Test, following the occurrence of the Fiordland event, and at the end of the testing period, the cumulative test results indicate that the EEPAS-0F model and the PPE model are both consistent with the data. Similarly, the EEPAS-1F and EEPAS-1R models are also consistent with the data in the cumulative L-Test; however, the EEPAS-0R model is rejected when tested over this time period.

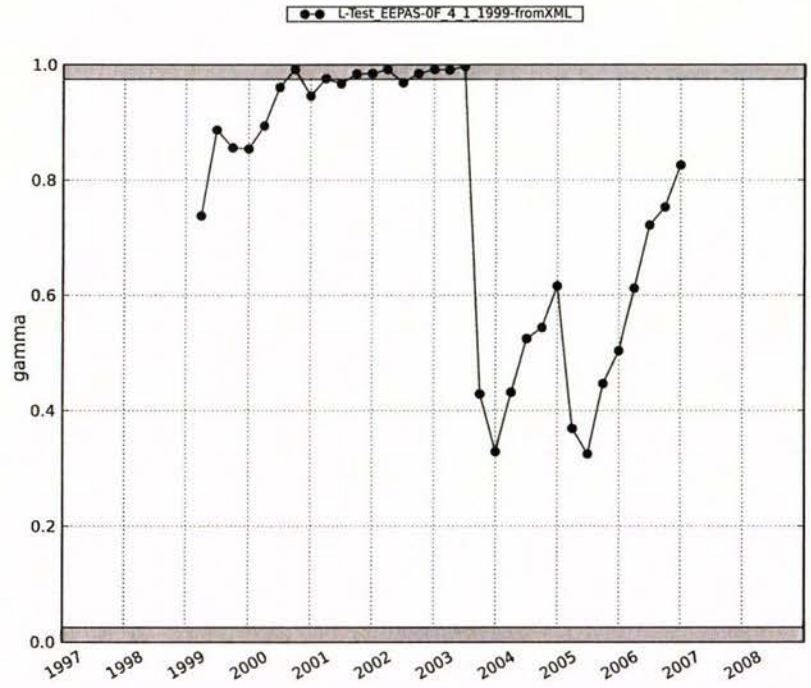

Figure 17 Cumulative L-Test results for the EEPAS-0F model for 1996-2007

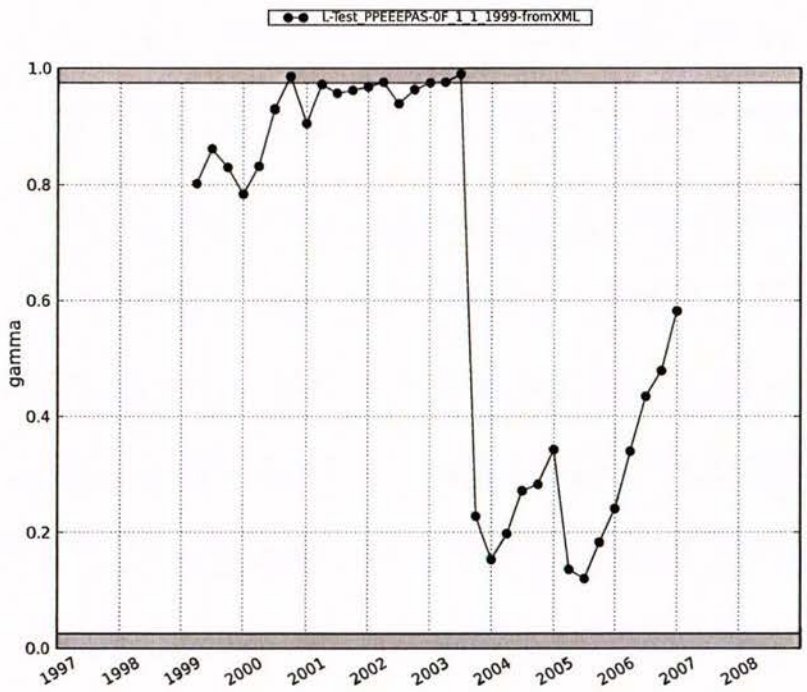

Figure 18 Cumulative L-Test results for the PPE model for 1996-2007.

5.5 Three-Month R-Test

The result of the cumulative R-Test comparing the EEPAS-0F model and the EEPAS-0R model is shown in Figure 19 and it can be seen that the EEPAS-0R model can be rejected based on the EEPAS-0F model (i.e., the EEPAS-0R forecast is significantly poorer). With five different models in this class, the R-Test involves 20 different model comparisons.

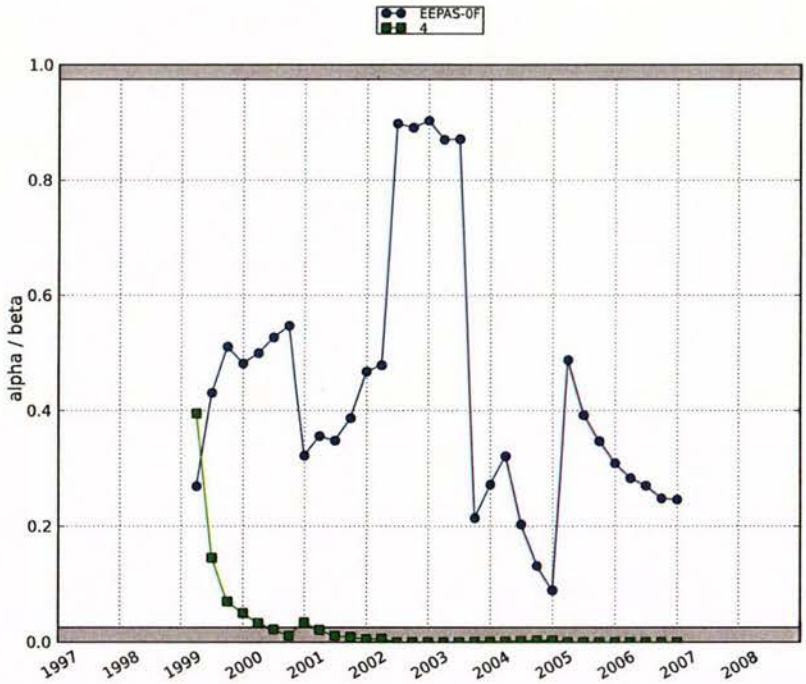

Figure 19 The cumulative R-Test results comparing the EEPAS-0F & EEPAS-0R models. The blue line represents the case when the EEPAS-0F model is considered the null hypothesis. In this test, the EEPAS-0R model is rejected.

Table 4 Summary of the results from all of the R-Test comparisons for the three-month models. R=Rejected; W=Not Rejected. The model shown in the top horizontal column is the model that was considered the null hypothesis for the test and the W or R refers to this model.

	Null Hypothesis				
	EEPAS-0F	EEPAS-0R	EEPAS-1F	EEPAS-1R	PPE
EEPAS-0F		R	R	R	R
EEPAS-0R	W		R	R	R
EEPAS-1F	W	R		R	R
EEPAS-1R	W	W	W		R
PPE	W	R	R	R	

A summary of all of the R-Test results is shown in Table 4. From these results it is clear that the EEPAS-0F model is the best performing model; all other models perform significantly poorer when compared to the EEPAS-0F model. The next best performing models are the EEPAS-0R model and the EEPAS-1F model which both perform significantly better than the EEPAS-1R model, but not than the PPE model. Neither the EEPAS-0R or the EEPAS-1F model is significantly distinguished from the other one. The next model in rank is the PPE model. While this model is rejected in all comparisons, in tests against the EEPAS-0R model and the EEPAS-1F model, the PPE model is indistinguishable from them. However, it is not possible to detect a significant difference between the performance of the PPE model and the EEPAS-1R model, which is significantly rejected when compared to all models except the PPE model (i.e., the EEPAS-1R model is significantly poorer than the others and the PPE cannot be distinguished from them).

In slightly more practical terms, the EEPAS-0F model is clearly significantly better than the others with only the EEPAS-0R and EEPAS-1F model showing any significant improvement over the remainder of the models.

5.6 One-Day Models

The one-day model class consists of models which aim to forecast aftershock activity and they put less emphasis on forecasting the main shock. Testing this class implies creating a forecast and evaluating it for every day in the testing period. As implemented within the testing centre testing two models, STEP and ETAS for a 10 year time period will take approximately 5 months to complete. One day takes about 1hr 20min to complete with 1 hour of this dedicated to the testing routines and the computational overhead. For this reason only the STEP and ETAS models were evaluated within the CSEP environment and have a limited result set at the time of reporting. To reduce the overhead for other evaluations and to greatly reduce the computation time, some tests were conducted using the CSEP tests, but outside of the environment. This reduced computation time to approximately 1 month. As these tests are optimisation exercises prior to submitting the models to the testing centre, it was not felt that the loss of rigor from not using the CSEP code was detrimental to the results. The testing done in this manner was comparison of the three variations of the STEP model discussed in the model classes section. It is hoped that the computational overhead will be reduced in future versions of the CSEP environment.

5.7 NZTC Tests

An example of the ETAS forecast generated for March 24th, 1996 is shown in Figure 20. One event greater than magnitude 4 occurred on this day within the white box on the map; this is an area of increased probability when compared to the surrounding area. An example STEP map, as calculated within the NZTC is show in Figure 21. Again, the event on this day occurred in an area of increased probability when compared to the surrounding area.

5.8 One-Day N-Test and L-Test Results

An error was discovered in the CSEP processing routines that calculate the cumulative scores from both the L-Test and N-Test. On days where zero events of magnitude greater than 4 occur, which are many, the models are always, and incorrectly rejected in these tests. It is in the CSEP plan to fix this bug and the forecasts will be re-evaluated in the future.

5.9 One-Day R-Test Results

The cumulative R-Test result for the ETAS and STEP models is shown in Figure 22. At the time of submitting the report, the results are for six months starting from January 1st, 2006. In the cumulative result, the ETAS model initially appeared to be producing significantly better forecasts than the STEP model, however by February the forecasts are statistically indistinguishable. As with some models in the five-year model class, each model may be rejected based upon the other model. This implies that each model contains some useful information that the other model does not. To best interpret these results will require continuation of the R-Test for a longer time period and evaluation of the results of the R-Test and L-Test after the CSEP code has been repaired.

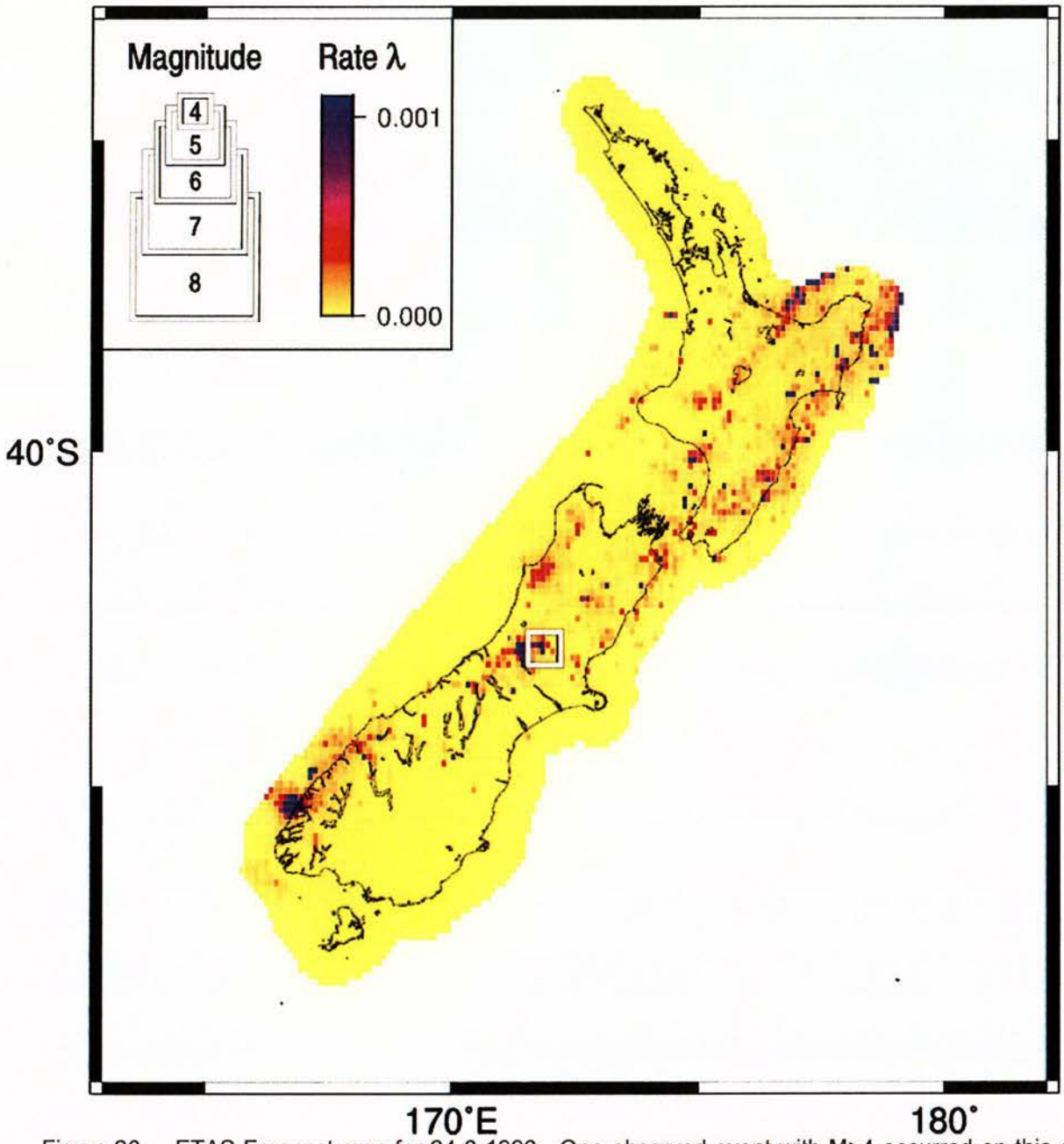

Figure 20 ETAS Forecast map for 24-3-1996. One observed event with M>4 occurred on this day and is show by the white box.

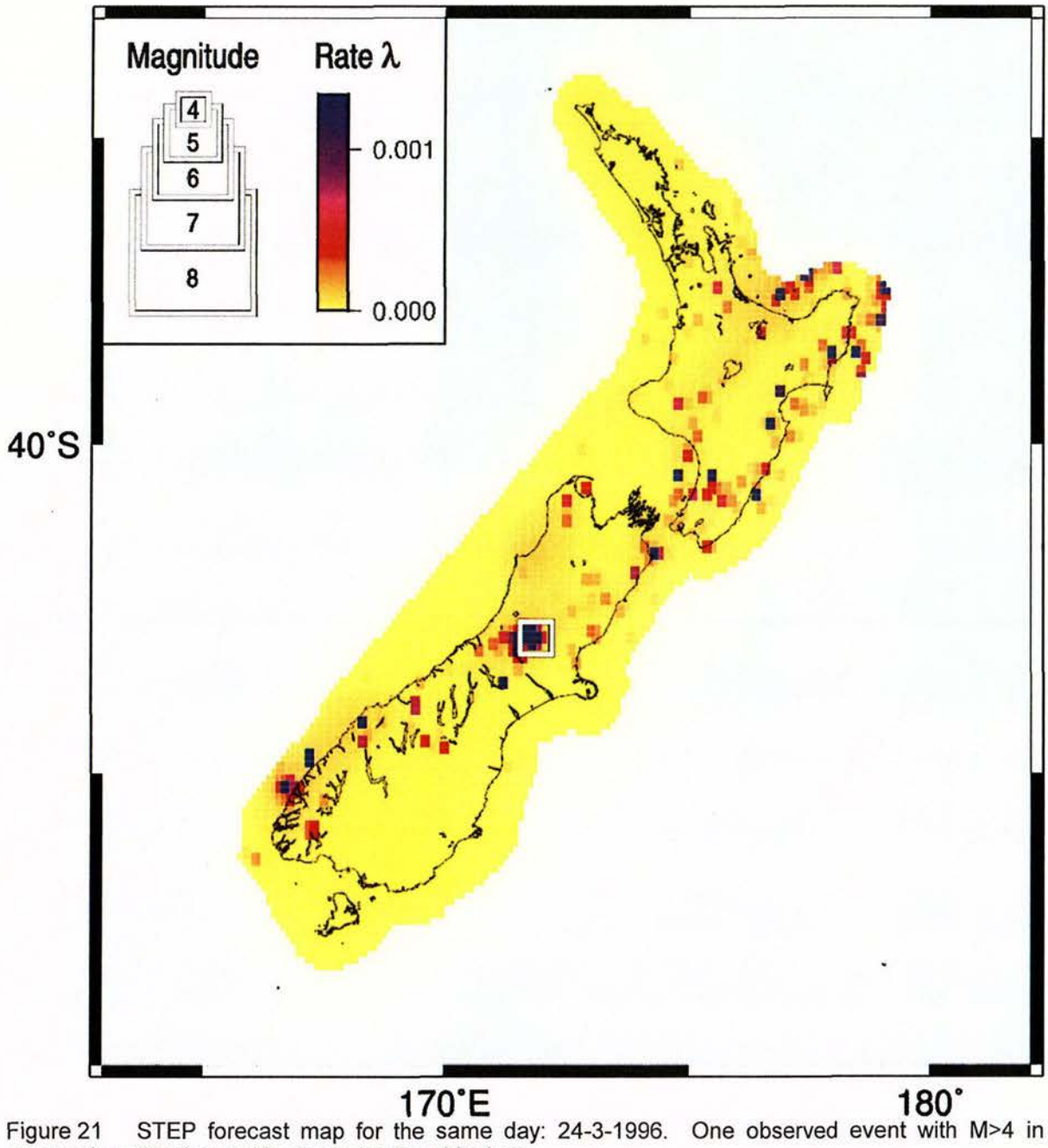

Figure 21 STEP forecast map for the same day: 24-3-1996. One observed event with M>4 in occurred on this day and is shown by the white box.

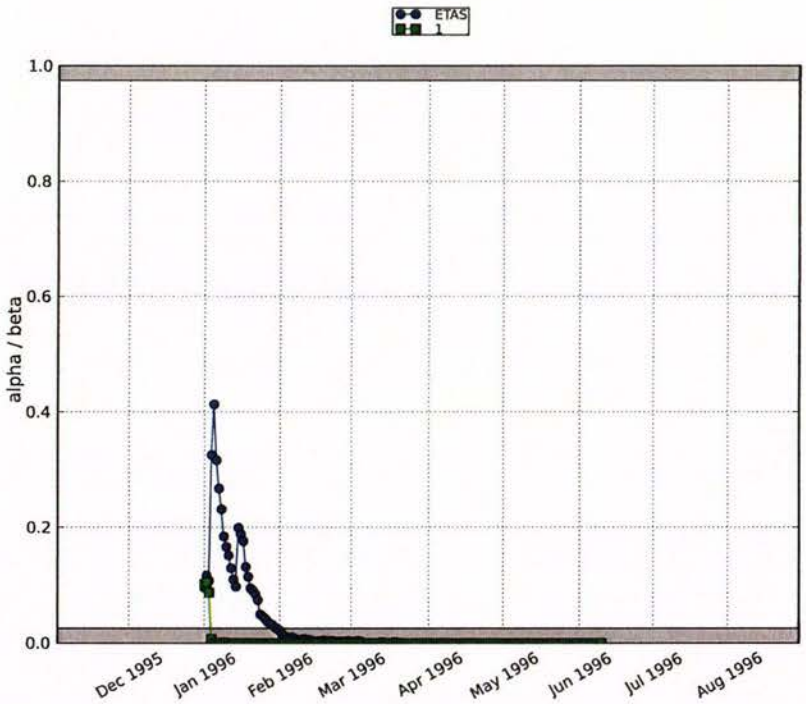

Figure 22 The Cumulative R-Test result for ETAS (blue) & STEP (green), January through June 1996

5.10 External One-Day Testing

To speed up the evaluation process we used exactly the same CSEP routines for the L-Test, N-Test and R-Test but operated them from a bare-minimum implementation in the MATLAB environment. In this case no archiving was done and minimal simulations were used. In this set-up we tested: 1) the original STEP implementation; 2) STEP using a generic model based on the Abundance model (Christophersen, 2005); and 3) a reformulation of the core Reasenberg and Jones (1989) model which is used by STEP (STEP-NG). The third model was formulated and implemented after initial testing of the STEP model revealed that a known mathematical error in the Reasenberg and Jones (1989) formulation resulted in significant errors in the New Zealand forecasts.

Figure 23 shows a comparison of the rates forecast from the original STEP implementation and STEP-NG. It is clear that the original STEP model greatly over-predicts the observed number of events. STEP-NG does a reasonable job of predicting the observed number of events and generally cannot be rejected in the L-Test and N-Test. The inset shows the forecast of the STEP-NG model in the days around the 2003 Fiordland event, which is shown prior to day 3,000 in the main plot. The inset shows the result of the failure of STEP-NG to forecast the Fiordland main shock, and hence its failure to predict the initial aftershocks in the remainder of the 24-hour testing period in which the Fiordland event occurred; on this day the STEP-NG model is rejected in the N-Test and the L-Test. This also demonstrates a drawback of only allowing the models to update once in any 24-time period as is required in the testing centre. Ideally a model would update with the occurrence of any "testable" event; this is a project for future improvement of the NZTC and CSEP testing code. This would also mean that models would only be evaluated when an event occurs which would greatly reduce the computational time required for one-day models.

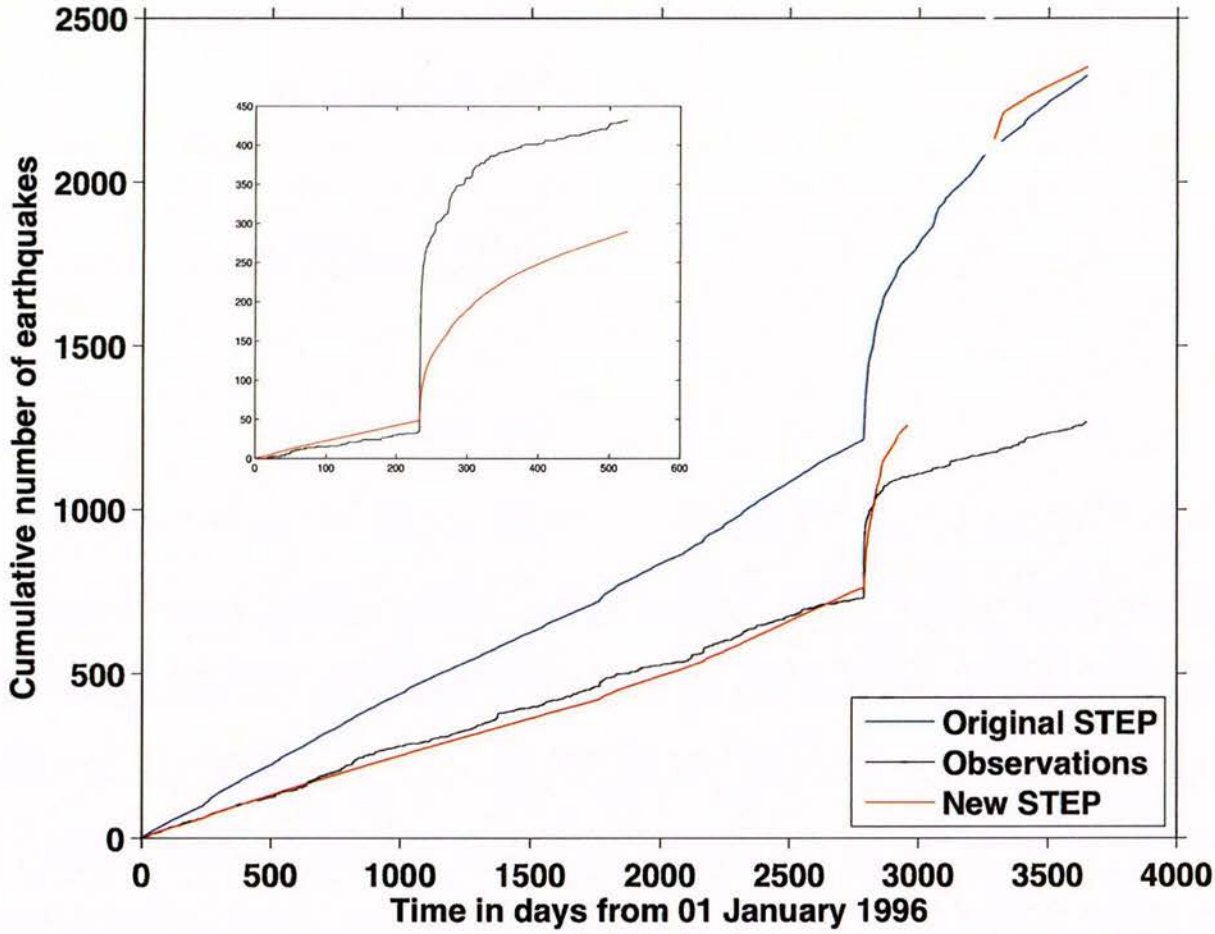

Figure 23 Comparison of the forecast rates of STEP, STEP-NG and the observations for 10 years starting in 1996. The inset shows the STEP-NG forecast and the observations following the 2003 Fiordland event; the offset from the missed mainshock and early aftershocks is apparent, but the rate after day 1 is similar.

6.0 DISCUSSION

As previously discussed, retrospective testing is not a replacement for rigorous prospective testing; however useful information can potentially be gained with retrospective tests. Poorly performing models can potentially be identified, and either removed from the extensive prospective testing procedures or corrected, if the poor performance is due to an error of implementation for CSEP testing; also model parameters may be better understood and optimised for better forecasting. Further, attempting to interpret a comprehensive set of standard CSEP test results can focus attention on how the presentation of the results could be improved.

A particular limitation of the testing presented here is in the five-year model class. The forecasts tested in this class are static, time-invariant forecasts based on the latest catalogue information up to the date of the creation of the forecast. For example, the NSHM-GB model used all earthquakes occurring between 1843 and 2006 for creating its forecast. The test of this model was done using earthquake data occurring between 1984 and 2009. The majority of these earthquakes were also used in developing the model, which creates a circular test. However this test still has value for two reasons: 1) if a model cannot explain the data that was used to create it, it clearly is a poor model; and 2) all three models tested in the 5-year

class used earthquakes from the testing period as a subset of the earthquakes used for developing the model, so a relative comparison also retains some value.

The most suitable time-length for testing the five-year models is a matter for ongoing consideration. As with any test: the longer the test, the more targeted earthquakes occur, and therefore the greater the power of the test becomes. However, the time-length becomes more critical for a model that is aimed at forecasting really long-term rates. For the NSHM-GB model, it may be a fair assumption that the forecast should not be used for a period of less than 50 years, the time period for which the models were designed. The 25-year period used is shorter than would be a fair test of the NSHM-GB, but a longer retrospective test period is prevented by the reduced quality of the catalogue in earlier times. It is possible that a longer time-period might produce different results, e.g., the NSHM-GB might be shown to be a significant improvement over the SUP model or PPE model if it is tested for 50 years. However, it is impossible to know without testing and it will likely be a function of whether or not future large earthquakes occur on mapped faults. At this point we feel that the 25 year period gives a fair representation of the performance of the long-term models. In order to best understand the performance of the NSHM, we will be supplementing the testing results from this study for the NSHM-GB with long-term ground-shaking based studies for the NSHM such as reported in Stirling and Gerstenberger (2009).

While the M8 model was rejected following the occurrence of the Fiordland event, it did show promise prior to the event. For this reason the minimal overhead required to test the M8 model in a prospective sense is justified so that we may gain a better understanding of the model and to see how it will handle future large events. It will also be valuable to test additional models in the six-month class. As a base, the SUP and PPE models will be implemented, but more can be learned by testing M8 against another time-varying model; we therefore intend to prospectively test the EEPAS model in the six-month class.

Prior to the retrospective testing it was noted by Rhoades et al. (2008) that the EEPAS-0F model was the optimal model for New Zealand in the fitting period and that the other versions of the EEPAS model did not fit the data as well. That the same result is confirmed independently by the CSEP testing centre is evidence that the EEPAS model has been correctly implemented for CSEP testing. Prospective testing can now be continued with some confidence.

In some measure, we have learned the least about the one-day model class due to the long-time period required to complete 10 years of testing and due to the errors in the cumulative testing of the L-Test and the R-Test. The tests are currently running continual calculations on the NZTC servers and we will allow them to run until the 10 year time period has completed. Then we will be able to evaluate the relative performance of the ETAS and the STEP model.

Despite the limited results available from the NZTC based testing for the one-day model class, we still gained important insight into the STEP model through the external testing. We have learned that the formulation of the core of the STEP model contains a mathematical error, and although it does not necessarily effect all implementations of the model, in New Zealand this error is critical. In the original formulation the scaling of the productivity of aftershocks with magnitude is proportional to the b-value of the Gutenberg-Richter relationship; this is a feature which has no physical basis but appears to work by coincidence in places such as California. In learning how poorly the model performed in New Zealand we

were directed to focus our efforts in reformulating the model and appear now to have a much improved model for prospective testing in New Zealand as well as testing centres in California and Europe.

As for the standard CSEP presentation of the testing results, it is clear that this could be improved in a number of ways. The present formats cover the matter of statistical significance well, but leave out more basic information, the inclusion of which would make the tests far more transparent. For example, neither the number of targeted earthquakes occurring during the test period nor the model (log-) likelihoods are included in standard presentation of the results. The absence of such information can make the results seem somewhat mysterious.

7.0 ALTERNATIVE TESTING METHODOLOGIES

7.1 Efficiency of earthquake likelihood model testing

The methods adopted by the earthquake forecast testing centres of the Collaboratory for the Study of Earthquake Predictability (CSEP) for real-time testing of earthquake forecasting methods involve the use of numerous synthetic earthquake catalogues generated from the expected number of earthquakes in each of many cells defined by time, magnitude and location to determine statistical significance in tests of each model. The tests adopted are the so-called *N*-test of the number of earthquakes predicted by the model, the *L*-test of the likelihood of the earthquake catalogue under the model, and the *R*-test of the relative likelihoods of two competing models (Schorlemmer et al., 2007). The efficiency of the calculations is an important issue, particularly for 24-hour forecasts which make the greatest demand on computer processing and storage, because the model forecasts must be updated and assessed daily. For the *R*-test, the computational burden is proportional to the square of the number of models being tested. Therefore, in order to be able to easily accommodate a large number of models in these tests, the procedures used should be as efficient as possible.

Broadly speaking, there are two issues here. The first is that when synthetic catalogues are generated, the most efficient method available should be used to do it. The second is that such simulation methods should be used only when necessary, and not when an equivalent exact test or a more efficient approximate test could be used. Therefore it is worth considering whether the intent of the tests presently implemented could be exactly or approximately accomplished more efficiently in some other way.

A regional earthquake likelihood model Λ is defined by the expected number of earthquakes λ_i , $i = 1, \ldots, n\}$ in each of a large number n of cells specified by limits in magnitude, location and time. These expected values are assumed to be the means of independent Poisson random variables (Schorlemmer et al., 2007). The log likelihood L of a catalogue Ω under the model Λ is then given by

$$L(\Omega \mid \Lambda) = \sum_{i=1}^{n} (\omega_i \ln\lambda_i - \lambda_i - \ln\omega_i!) \qquad (1)$$

where ω_i is the number of earthquakes occurring in the ith cell. Schorlemmer et al. (2007) described statistical tests to be carried out with the aid of synthetic earthquake catalogues

generated by simulating, in each cell, a Poisson random variable with the specified expected value. These tests are known as the *N*-test, *L*-test and *R*-test.

7.1.1 N-test

The purpose of the *N*-test is to compare the observed number of earthquakes in the test region with the number expected under each model. The testing framework implies independent Poisson distributions for the number of earthquakes in each cell. Therefore, the total number of earthquakes expected, $\hat{N}$, is the sum of the cell expected values,

$$\hat{N} = \sum_{i=1}^{n} \lambda_i \,. \tag{2}$$

Under the model, since the sum of a number of independent Poisson random variables is itself a Poisson random variable, the observed total number of earthquakes *N* is the realization a Poisson random variable with mean $\hat{N}$. The model is consistent with the data if $\hat{N}$ lies inside the Poisson confidence limits for the expected total number, given *N*. There is therefore no need to generate synthetic catalogues in order to execute the *N*-test. The "analytical" version of the test which does not involve synthetic catalogues is denoted here as the N_A-test.

A schematic of the NA test is shown in Figure 24. This is different in several ways from the standard CSEP presentation of the N-test. First, it shows the actual number of earthquakes observed during the trials. Secondly it presents the results for a number of models on the same plot. Thirdly, the significance tests just whether the model expectation lies inside the Confidence interval for the true Poisson expectation $\hat{N}$ given the observed number of earthquakes N. This confidence interval is computed from the observed N and Poisson distribution without reference to any model.

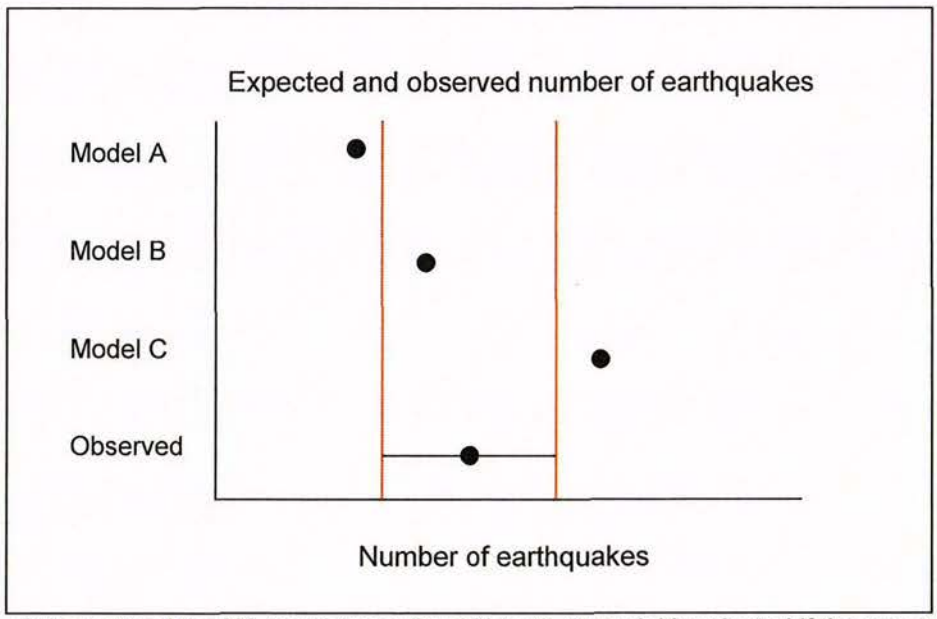

Figure 24 Schematic of the NA-test (streamlined N test). A model is rejected if the expected number of earthquakes under the model lies outside the 95% confidence limits for the mean of a Poisson distribution given the observed number of earthquakes N, as for models A and C in this diagram.

7.1.2 L-test

The *L*-test of Schorlemmer et al. (2007) compares the likelihood of the earthquake catalogue under the model with the likelihood of synthetic earthquake catalogues conforming to the model, in order to establish whether the observed likelihood is consistent with the model. In practice, the number of earthquakes is so small relative to the number of cells that $\omega_i = 0$ in the vast majority of cells, $\omega_i = 1$ in a few cells, and $\omega_i > 1$ very rarely.

Suppose that the *N* observed earthquakes occur in cells {*ik*, *k* = 1, ..., *N*}, where the numbers *ik* are not necessarily all different. The log likelihood in (1) can then be written as

$$L(\Omega \mid \Lambda) = \sum_{k=1}^{N} (\ln \lambda_{ik}) - \hat{N} - \sum_{k=1}^{N} (\ln \omega_{ik}!) \tag{3}$$

where λ_{ik}, i = 1,..., *N* are the expected numbers of earthquakes in cells in which earthquakes occur.

The third term is independent of the model and therefore of no account for comparing likelihoods of alternative models. It can be viewed as a penalty for discretization of continuous variables (magnitude, time, and location) into cells. It is zero except where two or more earthquakes occur in the same cell, which hardly ever happens. Therefore, for practical purposes, we can neglect the last term, and write simply

$$L(\Omega \mid \Lambda) = \sum_{k=1}^{N} (\ln \lambda_{ik}) - \hat{N}. \tag{4}$$

Let us now consider the uncertainty of $L(\Omega \mid \Lambda)$ before the catalogue is known. Under the model Λ, there is uncertainty both in the number of earthquakes *N* and the cells (*ik*, *k*=1, ..., *N*) into which they will fall. Given that there are *N* earthquakes, their log likelihoods are independent and identically distributed with cumulative distribution function $F_{\ln\Lambda}(\ln\lambda)$ determined by the cell expected values. If the cell expected values are ordered, so that $\lambda_{[1]} \le \lambda_{[2]} \le .... \le \lambda_{[n]}$, then

$$F_{\ln\Lambda}(\ln\lambda_{[k]}) = \frac{\sum_{i=1}^{k} \lambda_{[i]}}{\hat{N}}. \tag{5}$$

Figures 25 and 26 illustrate the difference between the distribution of cell expectations and the distribution of earthquake-cell expectations for the EEPAS model in one three month time period.

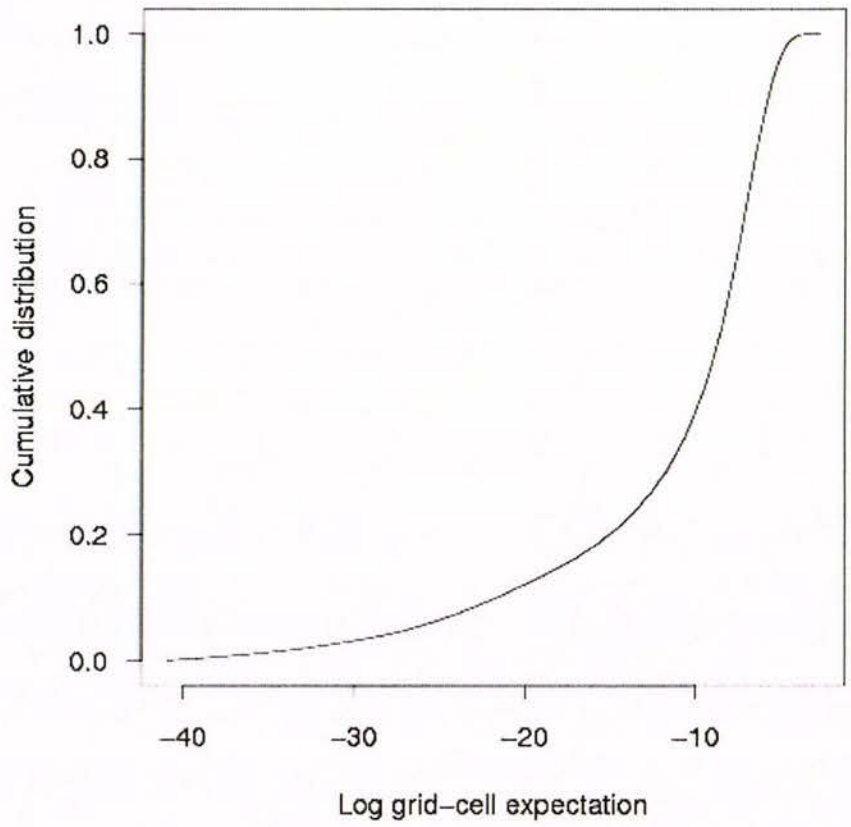

Figure 25 Distribution of grid-cell expectations for a three-month forecast using the EEPAS model.

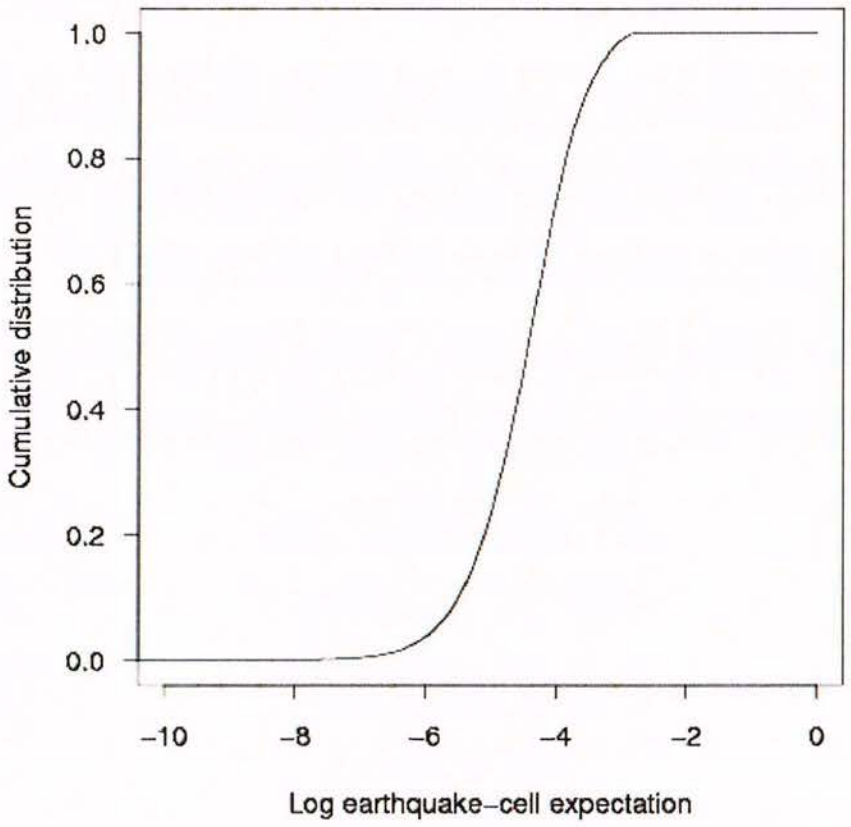

Figure 26 Distribution of earthquake-cell expectations for the same EEPAS model forecast as in Figure 25.

The log likelihood of an earthquake catalogue conforming to model Λ may be viewed as resulting from a two-step sampling procedure. First, the total number of earthquakes N is determined as the value of a Poisson random variable with mean $\hat{N}$. Secondly, the log likelihood of these N earthquakes is the sum of N independent random variables with distribution function $F_{\ln\Lambda}$. This sampling procedure provides an alternative means of simulating the log-likelihood statistics of catalogues consistent with the model Λ, which is far more efficient than the procedure defined by Schorlemmer et al. (2007) because it involves the simulation of only $N + 1$ random variables, rather than n.

Since the distribution of ln Λ is known exactly from the cell expectations, there is an alternative way of numerically approximating the distribution of the log-likelihood statistic. The cumulative distribution of L is represented exactly by

$$F_L(y-\hat{N}) = p(0 \mid \hat{N}) + \sum_{N=1}^{\infty}\left[p\left(N \mid \hat{N}\right) \sum_{\ln\lambda_{i1}+\cdots+\ln\lambda_{iN}<y} \left(\prod_{k=1}^{N} \frac{\lambda_{ik}}{\hat{N}} \right) \right] \tag{6}$$

where $p(N \mid \mu)$ is the Poisson probability of observing N events when the expected number is μ, and the summation inside the square brackets represents the distribution function of the sum of N independent random variables with cumulative distribution $F_{\ln\Lambda}$. Numerical approximation of F_L could be used to estimate the uncertainty of the log likelihood statistic and hence execute the L-test, as an alternative to generating synthetic catalogues, and the relative efficiency of these two approaches could be examined. The L-test obtained by such numerical approximation is denoted the L_N-test.

The summation inside the square brackets in (6) is the cumulative distribution of the conditional likelihood given the number of earthquakes N, under the model, and this distribution is of interest in its own right. This conditional, or *partial* likelihood, denoted L_P, has cumulative distribution function

$$F_{L_P}(y-\hat{N}) = \sum_{\ln\lambda_{i1}+\cdots+\ln\lambda_{iN}<y} \left(\prod_{k=1}^{N} \frac{\lambda_{ik}}{\hat{N}} \right). \tag{7}$$

A comparison of the observed likelihood with the distribution of L_P, called here the L_P-test, shows whether the earthquakes that do occur are consistent with the relative cell expectations in the model.

The distribution of the conditional likelihood could of course be estimated either from synthetic catalogues conforming to the model, with each catalogue containing exactly N earthquakes with log likelihoods conforming to the distribution $F_{\ln\Lambda}$, or from numerical approximation of the distribution of the sum of convolution of N identical distribution functions $F_{\ln\Lambda}(\ln\lambda)$. The L_P-test in conjunction with the N_A-test gives a rather complete description of the consistency of the model with the observed catalogue.

In a similar vein to the L_P test, we note that a standard test such as the Kolmogorov-Smirnov (K-S) test (Fisz, 1963, and other standard texts on mathematical statistics) could be used to compare the empirical distribution of $\{\ln\lambda_{ik},\ k = 1,\ldots,N\}$ with $F_{\ln\Lambda}$, as another test of whether the earthquakes that do occur are consistent with the relative likelihoods in the model. For large samples, i.e., large enough values of N, the Central Limit Theorem ensures that

$\sum_{k=1}^{N} \ln \lambda_{ik}$ is approximately normally distributed with mean $N\mu$ and variance $N\sigma^2$ where μ and σ are the mean and standard deviation, respectively, of ln Λ. Thus the cumulative distribution of $L(\Omega \mid \Lambda)$ may be written approximately as

$$F_L(x-\hat{N}) = p(0 \mid \hat{N}) + \sum_{N=1}^{\infty} p(N \mid \hat{N}) \Phi\left(\frac{x - N\mu}{\sqrt{N}\sigma}\right) \tag{8}$$

where Φ is the standard normal cumulative distribution function, and the distribution of L_P is given approximately by

$$F_{L_p}(x-\hat{N}) = \Phi\left(\frac{x - N\mu}{\sqrt{N}\sigma}\right). \tag{9}$$

How large N has to be to ensure a good approximation would depend on the distribution $F_{\ln\Lambda}(\ln\lambda)$, but for some models at least, even a moderate value of N might suffice. For example, in the EEPAS model example of Figures 25 and 26, this distribution is not wildly different from a normal distribution (Figure 27).

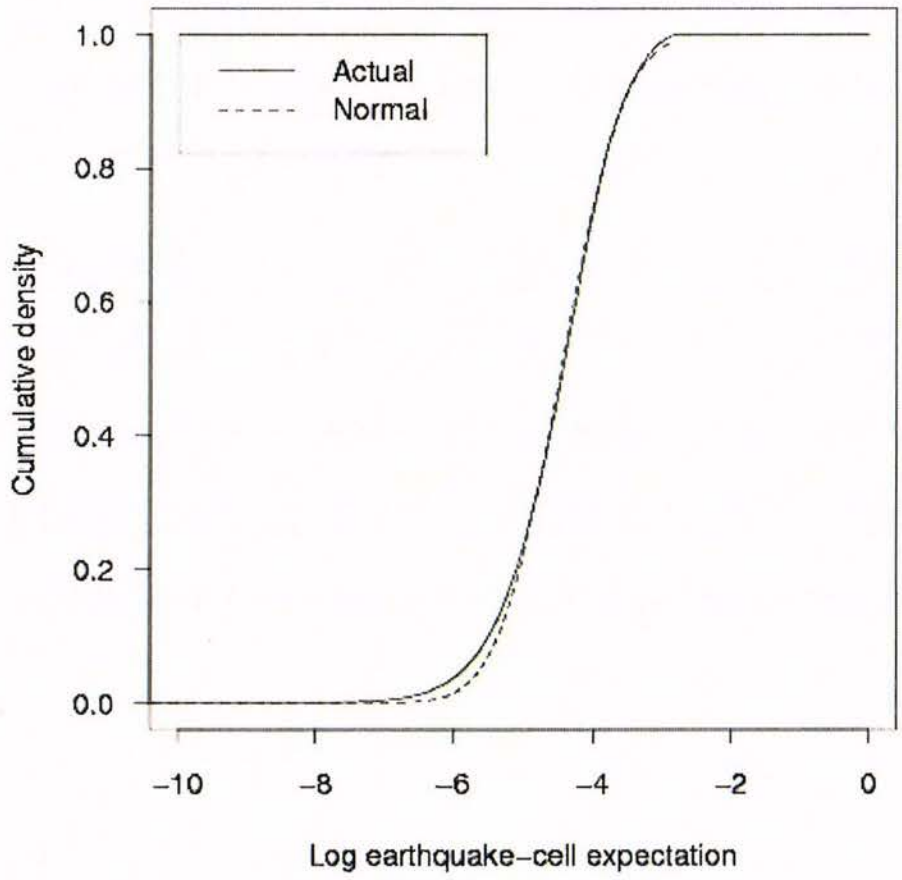

Figure 27 Distribution of log earthquake-cell expectation and fitted normal distribution.

Whatever form of L-test is used – the L-test as proposed by Schorlemmer et al. (2007), or L_N

or L_P as proposed here – a presentation according to schematic of Figure 28 has merit. Unlike the standard L-test results generated by the CSEP software, it shows the individual model likelihoods as well as combining the tests for a number of models into a single graphic.

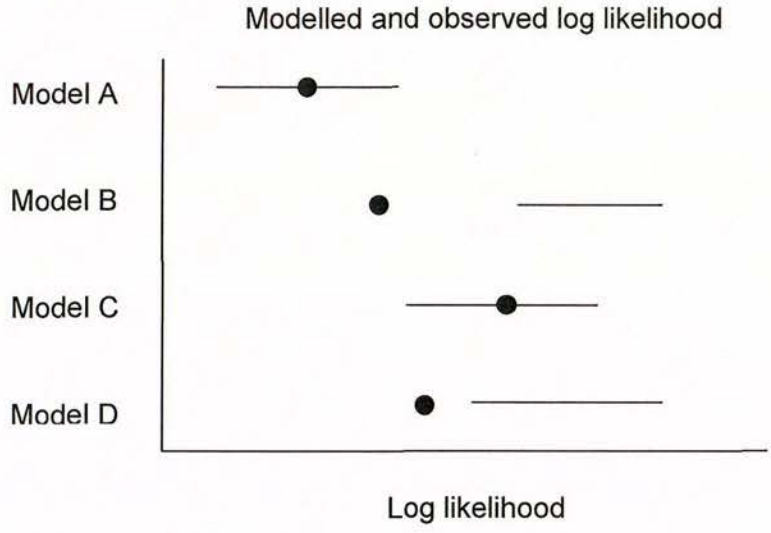

Figure 28 Schematic of L-test. A model is rejected if the observed likelihood (shown by a dot) lies outside the of the tolerance interval (horizontal line) for the likelihood under the model.

7.1.3 R-test

The purpose of the *R*-test is to compare the log likelihood of two different models Λ^1 and Λ^2, where one of the models, say Λ^2 is regarded as the null hypothesis. The test implemented in the testing centres is the one described by Schorlemmer et al. (2007) which involves evaluating the log likelihood ratio statistic $R = L(\Omega \mid \Lambda^1) - L(\Omega \mid \Lambda^2)$ for many synthetic catalogues conforming to Λ^2. If a small enough number of the synthetic *R* statistics are less than the value of *R* for the actual catalogue, the null hypothesis may be rejected.

Again it is worth considering whether there is a more efficient alternative than resorting to synthetic catalogue generation. Note that *R* can be written as

$$R = \sum_{k=1}^{N} \left(\ln \lambda_{ik}^1 - \ln \lambda_{ik}^2 \right) - \left(\hat{N}^1 - \hat{N}^2 \right), \tag{10}$$

and that, unlike equation (4), this expression is exact and does not depend on the assumption that no more than one earthquake occurs in any cell.

The distribution of Λ^2 is known. Also known is the distribution of Λ^1- Λ^2 for values of Λ^2 in any class interval, which can be approximated by a histogram with arbitrarily narrow class intervals (see Figure 29), showing the number of cells for which $\ln \lambda^1{}_i - \ln \lambda^2{}_i$ falls each class interval for $\ln\Lambda^1$- $\ln\Lambda^2$, given that $\ln \lambda^2{}_i$ falls into a certain class interval for Λ^2. Therefore the distribution of *R* can be approximated with arbitrarily small error by choosing narrow enough class intervals for $\ln\Lambda^2$ and ($\ln\Lambda^1 - \ln\Lambda^2$).

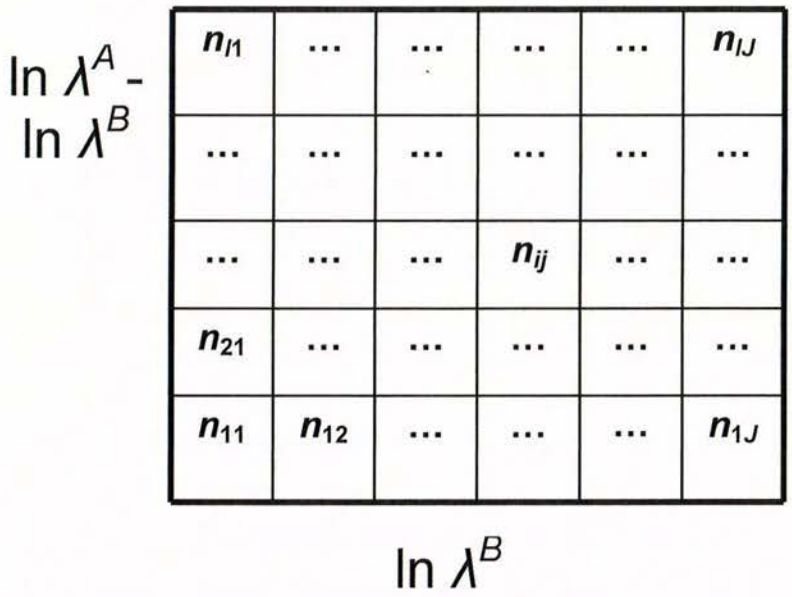

Figure 29 Schematic of data storage for approximating the distribution of the difference of the log-likelihood between models A and B assuming that model B is correct.

The alternatives available for the R-test are analogous to those available for the L-test. There is an R_N test, analogous to the L_N-test, which is equivalent to the existing R-test implemented in the CSEP testing centres, but achieved by direct numerical approximation of the distribution of R, rather than by simulation. There is an R_P-test, analogous to the L_P-test, in which is based on the distribution of R conditional on the number of earthquakes in the test period.

Again in this case, the Central Limit Theorem affords another means of approximating the R_P-test when the number of earthquakes in the test period is sufficiently large. If $\sum_{k=1}^{N} \ln \lambda_{ik}^{j}$ is normally distributed with mean $N\mu_j$ and variance $N\sigma_j^2$, the difference $\sum_{k=1}^{N} \left(\ln \lambda_{ik}^{1} - \ln \lambda_{ik}^{2} \right)$ is also normally distributed, with mean $N(\mu_1 - \mu_2)$ and variance $N(\sigma_1^2 + \sigma_2^2)$.

Again, whatever form of R-test is used, there is merit in adopting a presentation similar to Figure 30, in which both the difference between the log-likelihoods of the pairs of models, and a tolerance interval for this difference given each model are displayed. This figure shows the range of possibilities that can arise. Particular attention is drawn to the comparison of models A and C in this figure, in which model C fails to reject model A, but model A succeeds in rejecting model A, even though the log-likelihood is higher under model C. In this case, it is model C which is more informative, but the standard CSEP presentation of the R-test results would not show it.

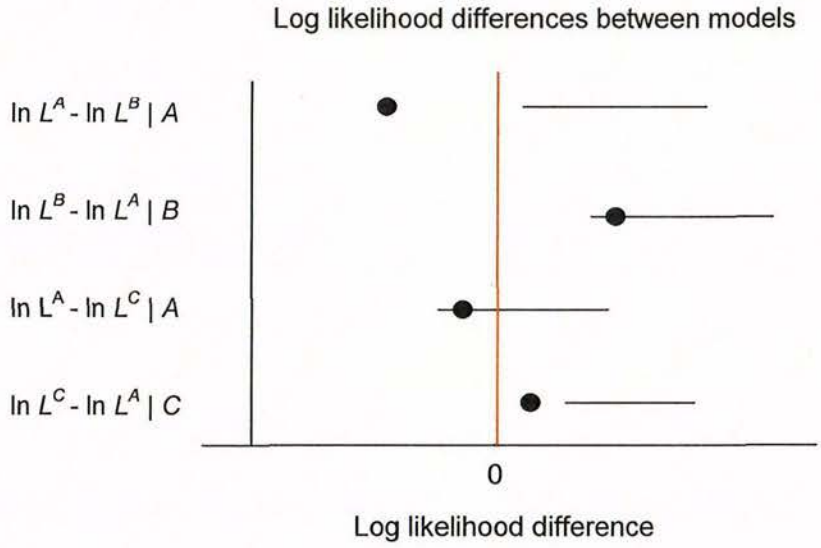

Figure 30 Schematic of proposed R-test presentation

7.1.4 Discussion

Although this analysis is primarily concerned with efficiency of the testing procedures, the more important issue is what test should be performed. In this regard, it is clear that core tests presently used by the testing centres are all sensitive to the number of earthquakes occurring during the test period, for the following reasons:

1. Schorlemmer et al. (2009) have shown that earthquake clustering is such that the total number of earthquakes occurring in a region during different time periods is more closely approximated by the negative binomial distribution than the Poisson distribution. This indicates that all models should at some stage fail the N-test based on the Poissonian assumption. This is not a reflection on the models, but only on the nature of earthquake clustering.
2. The log-likelihood statistic (4) is highly sensitive to the number of earthquakes N. This means that the L-test is also sensitive to N. It is obvious from a comparison of Figures 5 and 7 or 6 and 8 that the L-test and N-test results are highly correlated with each other.
3. Since the R-test involves a difference between two likelihood statistics, it too must be sensitive to N. A simple thought-experiment shows that if two models have different $\hat{N}$ values, but are otherwise similar, the one with the highest $\hat{N}$ value will eventually reject the other in an R-test, even if it overestimates the number of earthquakes and the alternative model doesn't.
4. When a model fails the L-test of R-test, it is impossible to tell from the standard CSEP testing output what the reason is for its failure, i.e., whether the failure is related to the number of earthquakes or the distribution of earthquake expectation over time, magnitude and location.

Therefore, a case can be made for removing the effects of the number of earthquakes from the L-test and R-test, since it is already been dealt with in isolation in the N-test. The core-test would then become the N_A-test, the L_P test and the R_P-test. These tests can be computed much more efficiently than the present tests.

7.2 Protocol for Testing Long-term Forecasts and Seismic Hazard Models

In the development of a testing centre for validation of probabilistic seismic hazard (PSH) models, careful consideration needs to be devoted to developing and appropriately administering the testing criteria. Short term forecast models (e.g. 10^0-10^1 years) can be tested on the basis of short-term observations of earthquake occurrence, but long term forecasts (>10^1 years) require longer catalogues that may include pre-instrumental data. Forecasts such as those used for building and planning developments (10^2-10^3 years) are typically developed from a combination of historical and geological datasets, and are frequently used to forecast for time periods much longer than those of any historical records. Ideally, tests for such long term hazard estimates should utilise pre-historical criteria, and this presents three obvious challenges: 1) identifying pre-historical criteria that can provide constraints on a component of a PSH model, or on the whole model. An example of the former would be to test predicted earthquake occurrence at a single site against a paleo-earthquake record for that site, and an example of the latter would be to test for consistency between predicted ground motions and the seismic fragility of ancient geomorphic features such as fractured cliff faces; 2) the lack of independence of the test and PSH model (i.e. a PSH model is likely to include the historical data in the model's construction) and; 3) the availability and increased uncertainties in the historical data.

The following is an outline of what methods, factors, parameters and variables should be considered when developing testing criteria for long-term PSH models.

7.2.1 Test Category

Because different tests target different information, different tests should be carried out on a PSH model, rather than basing all evaluations on just one test method. Two categories of test are reported here (with examples given for each):

7.2.2 PSH model component tests

Magnitude-frequency distribution: Is the historical catalogue within the range of possible synthetic catalogues produced by the source model?

Earthquake locations: Have earthquakes actually occurred in areas where they are predicted by the model for a time period equivalent to that of the historical period?

Moment rate budget: Does the sum moment rate for a region agree with the moment rate budget for that region from plate tectonics or (e.g.) GPS strain rate considerations?

Slip rate budget: Does the sum slip rate from one side of a region to the other agree with the plate tectonic motion rate for that region?

7.2.3 Whole PSH model tests

Is the predicted felt intensity or ground motion level for a given return period statistically indistinguishable from the observed felt intensity for a time period equivalent to that return period? Examples used in recent studies are MMI and instrumental PGA (Stirling & Petersen 2006; Stirling & Gerstenberger 2009).

Is the predicted ground motion level for a given return period less than or equal to the maximum ground motion constraints implied by the age and fragility of geomorphic features? Examples are studies of precariously-balanced rocks (e.g. Brune, et al., 2007) and fragile fractured cliff faces (e.g. Stirling et al., submitted).

Experience has shown us that different tests (such as those outlined above) can provide different answers in terms of evaluation of all, or part of, a PSH model. A testing protocol should include as many tests as possible, so that the model is evaluated in the most thorough way possible. A relevant example is where Stirling & Gerstenberger (2009) found that the 169 year historical MMI record-based testing of the New Zealand NSHM showed the model to under-predict the historical data, whereas a test based on more recent instrumental accelerogram data indicated consistency between model and data.

7.2.4 Relative Importance of Tests

In the above it is shown that multi-parameter testing requires the use of data that vary largely in terms of quality and quantity. Data issues contribute to the confidence we might have in a given test, and there can be important trade-offs between data quality and quantity, each of which may effect the power of the test. For example, tests that use the long non-instrumental record of felt intensities are high in quantity but low in quality, versus instrumental accelerogram-based tests which are high in quality and low in quantity. The results of these tests can be evaluated, compared and contrasted according to data quality and quantity. Ultimately, interpretation of the results can be complex and expert interpretation is necessary and becomes part of the overall process in a testing centre.

7.2.5 Testing Level

A significant decision to be made in setting up a testing protocol is what is a suitable and informative testing level for a given criteria. Some examples could be the lower magnitude threshold level for earthquake rate and area-based evaluation of a source model, and the lower MMI or PGA level for testing the output of a PSH model for a given return period. Magnitude 5 could be considered a worthwhile lower limit for rate and area-based testing, given that PSH models usually consider only earthquakes above magnitude 5. For lower earthquake ground-motion thresholds, recent studies (Stirling & Petersen 2006; Stirling & Gerstenberger 2009) used MMI 6 as a lower threshold test level as it provided a good trade off between shaking intensity and frequency (strong enough). Similarly a PGA level of 0.1g or greater gave a reasonably plentiful record of test data for damaging levels of ground- motion.

7.2.6 Testing period

To get statistically significant results, the testing period for a long-term PSH model should ideally be longer than the time span for which the model is forecasting. Long-term PSH models such as the NSHM are aimed at a wide range of time-periods and maybe be expected to forecast events with return periods of 10^2-10^3 years or longer in mind or they may be used to estimate hazard for structures with an expected life of 50 years; therefore the development of tests from prehistoric constraints (e.g. paleo-earthquake records and fragile geomorphic features) will greatly improve the value of testing when compared to tests based on catalogue data alone. Groundwork research on this topic area has been supported by EQC, Southern California Earthquake Centre (SCEC) and the US Department of Energy, but work still needs to be focussed on reducing uncertainties in age and fragility of the geomorphic features.

7.2.7 Retrospective versus Prospective Testing

In light of the above it is a near-impossible and impractical task to robustly test a long-term model in a prospective sense, but retrospective testing can provide confirmation that the model is consistent with what has happened in the past. As described above, any retrospective test is subject to bias, but it is also the simplest type of testing that any model must be able to pass to be considered acceptable.

8.0 CONCLUSIONS

The M8 algorithm, as originally formulated, was not at all suited to CSEP testing. However, following several significant modifications, a version of the M8 model has now been successfully implemented in the NZTC. It is the only six-month model currently installed in the testing centre.

Retrospective testing is not a replacement for rigorous prospective testing; but information from the retrospective testing of the models in the NZEFTC has nevertheless proved useful.

For the five-year models, the fact that the SUP, PPE, and NSHM-GB all pass the N-test and the L-test gives confidence that these models have been correctly implemented in the testing centre. The fact that the NSHM-GB is rejected by the PPE model in a 25-year retrospective R-test does not necessarily mean that the same result should be expected in prospective testing; this result is not at all surprising, given that the earthquakes from the testing period were used directly to build the PPE forecast. However, the rejection of the NSHM-GB model by the SUP model in a similar test is surprising, because the SUP model contains no information on the spatial distribution of earthquakes. It is probably due to the lack of earthquakes on mapped faults during the testing period. The most suitable time-length for testing of the NSHM-GB model is a matter of ongoing consideration. Additionally, it should be noted that this testing is not on a direct interpretation of the NSHM and the fault based information is different between the two models. At this point the reason for the PPE model to be rejected by the SUP model is not clear and will require further investigation.

For the 6-month model, M8, its rejection by the N-test and L-test over the period 1996-2007 was a result of its failure to forecast the numerous Fiordland earthquakes of 2003 and 2005. A more complete assessment of the M8 model can be made when other models, such as SUP, PPE and EEPAS, are added to the six-month class.

For the three-month models, that all models passed the N-test and all but one model, EEPAS_0R, passed the L-test was generally consistent with expectations. Also the fact that the EEPAS_0F model emerged as the best model in the retrospective test was consistent with what was already known from the fitting of these models, adds confidence that the EEPAS model is correctly installed for testing. It should not necessarily be expected that the same model will prove to be the best in prospective testing.

The tests of the one-day models have been affected by errors in the testing centre software, and lengthened by the computational overheads of the CSEP testing system, but when the ten-year retrospective tests of STEP and ETAS have been completed, will be useful in gauging the relative merits of these two models. Important insight into the STEP model has been gained through external testing. Identification of a mathematical error in its formulation, which particularly affects its performance in New Zealand, has resulted in the definition of a

much improved model for prospective testing in New Zealand as well as the testing centres in California and Europe.

The retrospective tests would be easier to interpret if the standard CSEP presentation of results were to include the number of targeted earthquakes and likelihood of the data under each model in the testing period.

We have proposed alternative statistical methods to accomplish the intent of the N-test, L-test and R-test. The alternative methods do not require time-consuming catalogue simulations, but in some cases require numerical approximation to known exact distributions. The tests can be implemented with much greater computational efficiency than the existing CSEP-implemented tests. If processing-capacity becomes an issue as the number of models being tested increases, the use of the proposed methods instead of the simulation-based tests could be advantageous. In any case, there is a need to reconsider what combinations of tests are provided by the testing centres, because the present tests are all sensitive to the number of earthquakes occurring during the test period.

9.0 ACKNOWLEDGEMENTS

This work was funded by the EQC Research Foundation under project number 08/557. Software developed by CSEP has been extensively used in this project and has been implemented with considerable help from M. Liukis and J. Yu. W. Smith and R. Robinson have provided helpful reviews of the report.

10.0 REFERENCES

Brownrigg, R. & Harte, D. (2005). Using R for statistical seismology. R News 5(1), 31–35. ISSN 1609-3631.

Brune, J. N., M. D. Purvance and A. Anooshehpoor, 2007, Gauging Earthquake Hazards with Precariously Balanced Rocks, *Am. Sci.*, 95, 36-43.

Christophersen, A. (2005) Towards a New Zealand model for short-term earthquake probability: Aftershock productivity and parameters from global catalogue analysis, Second progress report, EQC Project 6OPR1b.

Fisz, M., 1963, Probability Theory and Mathematical Statistics, Third edition, Wiley, New York, 677p.

Gerstenberger, M., 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. 2009, Time Varying Hazards Research Programme Number TVH2/11 Report, GNS Science Consultancy Report 2009/28.

Harte, D., Feng-Dong, L., Vreede, M., Vere-Jones, D., Wang, Q., 2007, Quantifying the M8 algorithm: model, forecast, and evaluation, *New Zealand Journal of Geology & Geophysics* **50**, 117-130.

Keilis-Borok VI, Kossobokov VG 1990. Premonitory activation of earthquake flow: algorithm M8. Physics of the Earth and Planetary Interiors 61: 73–83.

Reasenberg P.A. (1985). Second-order moment of central California seismicity, 1969-1982. *Journal of Geophysical Research* **90**, 5479–5496.

Reasenberg, P.A., and Jones, L.M., 1989, Earthquake Hazard After a Mainshock in California, *Science*, Vol. 243. no. 4895, pp. 1173 - 1176

Rhoades, D.A., Gerstenberger, M.C., Christophersen, A., Savage, M. ... Testing and Development of Earthquake Forecasting Models. EQC Research Project 06 GNS Science Consultancy Report 2008/70

Rhoades, D. A., and F. F. Evison (2004). Long-range earthquake forecasting with every earthquake a precursor according to scale. *Pure and Applied Geophysics* **161**, 47–71.

Robinson, R., and Benites, 1996, R., "Synthetic Seismicity Models for the Wellington Region, New Zealand: Implications for the Temporal Distribution of Large Events", J. Geophys. Res., 101, 27,833-27,845.

Schorlemmer, D.; Gerstenberger, M.C.; Wiemer, S.; Jackson, D.D.; Rhoades, D.A. 2007 Earthquake likelihood model testing. *Seismological Research Letters* **78** (1): 17-29

Schorlemmer, D., and Gerstenberger, M.C., 2007, RELM Testing Center, *Seismological Research Letters* **78** (1), 30-36.

Schorlemmer, D., Zechar, J.D., Werner, M.J., Field, E.H., Jackson, D.D., Jordan, T.H. and the RELM Working Group (2009), First results of the Regional Earthquake Likelihood Models experiment, Pure and Applied Geophysics, Seismogenesis and Earthquake Forecasting: the Frank Evison Volume, accepted for publication.

Stirling, M.W., McVerry, G.H., and Berryman, K.R. (2002). A new seismic hazard model for New Zealand. *Bull. Seismol. Soc. Amer.* **92**, 1878–1903.2.0

Stirling, M.W., and Petersen, M.D., 2006, Comparison of the historical record of earthquake hazard with the seismic hazard models for New Zealand and the continental United States, *Bulletin of the Seismological Society of America*, 96, 1978-1994.

Stirling, M. W., and Gerstenberger, M.C., (2009), Development of tests for long-term earthquake ground motion forecasts in New Zealand, EQC Research Project, GNS Science Consultancy Report 2009/140.

www.gns.cri.nz

Principal Location

1 Fairway Drive
Avalon
Lower Hutt 5010
PO Box 30368
Lower Hutt 5040
New Zealand
T +64-4-570 1444
F +64-4-570 4600

Other Locations

Dunedin Research Centre
764 Cumberland Street
Dunedin 9016
Private Bag 1930
Dunedin 9054
New Zealand
T +64-3-477 4050
F +64-3-477 5232

Wairakei Research Centre
114 Karetoto Road, Wairakei
Taupo 3377
Private Bag 2000
Taupo 3352
New Zealand
T +64-7-374 8211
F +64-7-374 8199

National Isotope Centre
30 Gracefield Road, Gracefield
Lower Hutt 5010
PO Box 31312
Lower Hutt 5040
New Zealand
T +64-4-570 1444
F +64-4-570 4657