CSEP Testing Center and the first results of the earthquake forecast testing experiment in Japan

H. Tsuruoka¹, N. Hirata¹, D. Schorlemmer^{3,2}, F. Euchner⁴, K. Z. Nanjo^{5,1}, and T. H. Jordan²

¹Earthquake Research Institute, University of Tokyo, Tokyo, Japan

²Southern California Earthquake Center, University of Southern California, Los Angeles, California, USA

³GeoForschungsZentrum, Potsdam, Germany

⁴Swiss Seismological Service, ETH Zürich, Zürich, Switzerland

⁵National Research Institute for Earth Science and Disaster Prevention, Tsukuba, Japan

(Received February 4, 2011; Revised May 7, 2012; Accepted June 20, 2012; Online published August 27, 2012)

Major objectives of the Japanese earthquake prediction research program for the period 2009–2013 are to create earthquake forecasting models and begin the prospective testing of these models against recorded seismicity. For this purpose, the Earthquake Research Institute of the University of Tokyo has joined an international partnership to create a Collaboratory for the Study of Earthquake Predictability (CSEP). Here, we describe a new infrastructure for developing and evaluating forecasting models—the CSEP Japan Testing Center—as well as some preliminary testing results. On 1 November 2009, the Testing Center started a prospective and competitive earthquake predictability experiment using the seismically active and well-instrumented region of Japan as a natural laboratory.

Key words: Earthquake forecasting and testing, statistical seismology, seismic hazard, Japan.

1. Introduction

Japanese research projects aimed of scientific earthquake prediction have primarily focused on a better understanding of the mechanism of earthquake occurrence and the development of predictive simulation technologies based on physical modeling of earthquakes (Hirata, 2004). Less emphasis has been placed on the creation and prospective testing of earthquake forecasting models. However, this objective has recently become a key challenge under the national “Observation and Research Program for Prediction of Earthquake and Volcanic Eruption (2009–2013)” (Hirata, 2009). Together with other subprograms associated with predictive simulation analysis of crustal dynamics, the project on “Earthquake Forecast System based on the Seismicity of Japan” has begun research on this task.

A key component of this project is the Japan Testing Center, a unique cyber infrastructure for the development and testing of stationary and time-varying seismicity forecasting models for Japan. This is one of several testing centers of the Collaboratory for the Study of Earthquake Predictability (CSEP), a global project for earthquake predictability research (Jordan, 2006). CSEP is the successor of the Regional Earthquake Likelihood Model (RELM) project (Field, 2007; Schorlemmer and Gerstenberger, 2007; Schorlemmer *et al.*, 2007). RELM aimed at generating a suite of earthquake forecast models for California and testing them rigorously against future ob-

servations. The Earthquake Research Institute (ERI) at the University of Tokyo joined CSEP to install the Japan Testing Center in summer 2008 through an international collaboration. As a test run, a first set of three one-year smoothed-seismicity models were fully implemented in the Testing Center starting 1 September 2008 (Tsuruoka *et al.*, 2008). On 1 November 2009, several prospective and competitive earthquake predictability experiments began in the testing region of Japan (Nanjo *et al.*, 2011). In this paper, we describe the Japan Testing Center that hosts these experiments and present the first experimental results. Because these results are based on only 3 months of observation, further testing in the same controlled environment will be needed to reach useful and meaningful results concerning the reliability and skill of the submitted earthquake forecasting methods. Comprehensive appraisal, and their implications on seismic hazard, of the different models will be discussed in future publications (Yokoi *et al.*, in preparation).

2. Japan Testing Center

One of the primary steps in launching an earthquake forecast experiment is to obtain a consensus among potential participants (Nanjo *et al.*, 2011). At an international symposium held in May 2009 at ERI, the attendees decided to employ all applicable rules that had been previously defined for other CSEP experiments such as those, for example, for California. The Japan Testing Center completely follows the technical design of CSEP Testing Centers that have been set up in California (Schorlemmer and Gerstenberger, 2007; Schorlemmer *et al.*, 2007, 2010b), Europe (Schorlemmer *et al.*, 2010a), and New Zealand (Gerstenberger and Rhoades, 2010). A computer system running the CSEP Testing Cen-

ter software is described by Zechar *et al.* (2010b).

The CSEP Testing Center software implements a number of statistical tests for model performance. For the current Japan Testing Center, the test suite consists of the L-, M-, N-, S-, and T/W-tests (Schorlemmer *et al.*, 2007, 2010b; Zechar *et al.*, 2010a; Rhoades *et al.*, 2011): the L-, M-, N-, and S-tests test the consistency between the observation and the forecast in the joint log-likelihood, the magnitude distribution, the total number, and the spatial distribution, respectively. As described by Schorlemmer *et al.* (2007) and Zechar *et al.* (2010a), the test results are given as quantile scores (δ_1 and δ_2 for the N-test, γ for the L-test, ζ for the S-test, κ for the M-test) of the observed value compared to simulated values based on the forecast. When δ_1 is very small, the forecast rate is too low (an underprediction); and, when δ_2 is very small, the forecast rate is too high (an overprediction). Following Schorlemmer *et al.* (2007, 2010b) and Zechar *et al.* (2010a), we use a two-sided significance level of 5%, where applicable, or a single-sided significance level of 2.5%. The T/W-test, the set of the T- and W-tests, is used to compare two forecasts, based on information gain per earthquake (Rhoades *et al.*, 2011). One model differs significantly from its competing models if either one or both of the T- and W-tests show significance at a 95% confidence level.

The test suite also includes the R(atio)-test that compares two forecasts (Schorlemmer *et al.*, 2007, 2010b), an alternative to the T/W-test. However, we do not include its result in our paper because several researchers working with other Testing Centers pointed out the difficulty of interpreting the R-test results in a straightforward fashion. For example, Gerstenberger *et al.* (2009) found that models sometimes mutually reject each other. We also faced a similar difficulty. The T/W-test is easier to interpret than the R-test.

The particular interest of this Testing Center are the experiments in the testing region of Japan, one of the most seismically active and well-instrumented regions in the world. To make full use of this location, rules and infrastructure for this experiment have been set up. Some specification to Japan is given below. For more details, visit the project website (<http://www.eic.eri.u-tokyo.ac.jp/ZISINyosoku/>).

2.1 Testing regions and testing classes

The selection of testing regions and testing classes, describing the rules for an experiment, has to be driven by a consensus process and is thus based on a wide range of requirements and/or limitation of forecast models and the research direction of potential participants. On the technical side, a full characterization of the earthquake catalog is necessary to ensure that the targeted magnitude range of an experiment will be fully recorded by the seismic network. In 2009, we performed such a characterization using the method of Schorlemmer and Woessner (2008), and we defined the testing region using criteria similar to Schorlemmer *et al.* (2010a) but also by taking into account the availability of models, the sufficiency of the input data, and potential applicability of forecasting results to the practical problems of risk mitigation. Taking all these considerations into account, we proposed 3 testing regions with 4 testing classes each, resulting in 12 experiments covering

different space and time scales:

Testing regions

- “All Japan” that covers the whole territory of Japan including about 100 km offshore and down to a depth of 100 km with a node spacing of 0.1° .
- “Mainland” that covers Japan’s mainland excluding any offshore areas and down to a depth of 30 km with a node spacing of 0.1° .
- “Kanto” that covers the Kanto district of Japan down to a depth of 100 km with a high-resolution node spacing of 0.05° .

Testing classes

- 1-day forecast: Forecast models must define earthquake rates for each magnitude bin in the magnitude range $4.0 \leq M \leq 9.0$ (0.1 magnitude unit steps) at each node for consecutive 1-day time windows, each starting at midnight UTC. The magnitude bin $M = 4.0$ covers the magnitude range $3.95 \leq M < 4.05$. The first forecast time window starts at midnight of 1 November 2009.
- 3-month forecast: Same as for the 1-day class but the forecast time-window length is 3 months. The first time window starts at midnight on 1 November 2009, the second time window starts at midnight of 1 February 2010, and so on.
- 1-year forecast: Same as for the 1-day class but the time-window length is 1 year and the magnitude range is $5.0 \leq M \leq 9.0$.
- 3-year forecast: Same as for the 1-year class but the time-window length is 3 years.

The rationale behind introducing these regions and classes is discussed in Nanjo *et al.* (2011).

2.2 Earthquake catalog

As part of the rules of each experiment, we have to define the earthquake catalog against which the forecasts will be tested (testing catalog) as well as the earthquake catalog that can be used as input data for generating earthquake forecasts (learning catalog). For all Japanese experiments, we chose to use the Japan Meteorological Agency (JMA) catalog covering Japan and surrounding areas. Because CSEP does not test earthquake forecast models against real-time earthquake data (which often contain erroneous locations), testing was performed after a delay that allows for manual revisions of earthquake locations by the network operators. This delay ensures a stable catalog for testing. In Japan, the delay is set to 6 months but can be extended if necessary due to seismic crises.

The JMA catalog includes earthquakes since 1923. In 2003, JMA started relocating past events and introduced the current magnitude scale (Japan Meteorological Agency, 2003; Nanjo *et al.*, 2010a). The recomputation of magnitudes for past events is not yet finished for the time prior to 1965. Therefore, we decided to use only data after 1965 for all experiments. Neither the testing catalog nor the learning catalog are declustered. In this paper, we use the testing catalog beginning at the submission deadline of 1 November 2009.

2.3 Requirement of model submission

The goal of CSEP is to facilitate earthquake forecast testing that is completely transparent, fully reproducible, and truly prospective. To meet these criteria, most CSEP experiments have imposed two conditions: (a) a forecasting model must be submitted as a source code, and (b) any updating of a forecast after the initiation of an experiment can only be done by running the code on a CSEP computer without any interaction with the authors of the model.

Criterion (a) ensures full documentation and reproducibility of the models. Exceptions were the initial 5-year RELM experiment (Field, 2007) and similar 5- and 10-year CSEP-Italy experiments (Marzocchi *et al.*, 2010) in which forecasts were registered as numerical tables rather than source codes, and authors documented their forecasts in papers submitted for publication prior to the experiment.

Criterion (b) ensures that the experimental evaluation is “blind” to the model authors. For forecasts that need to be updated during the experimental phase, the submission of an executable code is indispensable to blind prospective testing. Another practical reason is the catalog latency described above. Before submitting models to the Testing Center, modelers were allowed to use the JMA catalog as an input to their model development and optimization. However, to ensure that the experiments remained blind to the forecast authors, the evaluations were lagged by a time equal to the catalog latency.

3. Seismic Network Coverage and Earthquake Detection Capability

The earthquake detection capability of the seismic network is not only important for delineating the prospective testing region, but also for calibrating the statistical properties of seismicity during the learning period. Such calibrations can only be meaningful above the completeness threshold of the catalog, M_c , defined as the magnitude above which all earthquakes can be detected by the network. For example, studies of earthquake-size distributions and seismicity rates are highly dependent on these thresholds (e.g. Wiemer and Wyss, 2002; Schorlemmer *et al.*, 2005). We performed a retrospective analysis of recording completeness to provide participants with detailed knowledge about the input data (Nanjo *et al.*, 2010a).

To complement its own nationwide network, JMA started, in October 1997, real-time processing of waveform data from many other networks operated by Japanese universities and institutions. Among the non-JMA networks is Hi-net, a borehole seismic network of about 700 stations deployed by the National Research Institute for Earth Science and Disaster Prevention (Obara *et al.*, 2005). Nowadays, about 1200 seismometers are operating under the hybrid network and are detecting more than 100,000 events annually.

Nanjo *et al.* (2010a) conducted a comprehensive analysis of M_c in Japan using the JMA catalog. They computed M_c based on the Gutenberg-Richter frequency-magnitude relation (Gutenberg and Richter, 1944). Starting in the 1970s, completeness magnitudes were $M_c \geq 3$ and decreased significantly over the last decades due to increased coverage and density of the network. For the last few years, the com-

pleteness level of the JMA catalog is $M_c = 1.9$ for mainland Japan but does exhibit higher values offshore. Computations using the probability-based magnitude of completeness (PMC) method (Schorlemmer and Woessner, 2008; Nanjo *et al.*, 2010b), as used for the completeness assessment for the testing catalog, is underway (Schorlemmer *et al.*, 2008) and will provide additional and more detailed information on completeness.

4. First Results of the Earthquake Forecast Testing Experiment

We show the first test results to illustrate some experiments and to give a feeling of test performances. Because the experiments started only recently, when the original version of the manuscript was submitted on 4 February 2011, only the first round in all 3-month classes for the three different testing regions are available.

4.1 Testing classes and regions

As described in the previous section, the 3-month testing-class forecasts the rate, λ (number / 3 months), of earthquakes at each spatial cell in one of the three testing regions and for each magnitude bin in the range $4.0 \leq M \leq 9.0$ (in 0.1 magnitude unit steps). The three testing regions are: (a) “All Japan,” (b) “Mainland,” and (c) “Kanto” in Figs. 1, 2, and 3, respectively.

4.2 Earthquake forecast models

The models submitted to the 3-month class were built on different earthquake generation hypotheses. All models used past and current seismicity to extrapolate into future earthquake rates. All software-codes were submitted to the Testing Center before the start of the testing experiment. For the first 3-month period 1 November 2009–31 January 2010, the catalog with final solutions before 1 November 2009 was provided to the models by the Testing Center as the input dataset. For the “All Japan” region (Fig. 1), 9 models were submitted: HIST-ETAS5pa and HIST-ETAS7pa (Ogata, 2011), MARFS and MARFSTA (Smyth and Mori, 2011), Triple-S-Japan (Zecher and Jordan, 2010), and RI10k, R30k, RI50k, and RI100k (Nanjo, 2011). The 9 models for “Mainland” (Fig. 2) are EEPAS and PPE (Rhoades, 2011), MARFS, MARFSTA, Triple-S-Japan, RI10k, R30k, RI50k, and RI100k. The 7 “Kanto” models are (Fig. 3) HIST-ETAS5pa, HIST-ETAS7pa, Triple-S-Japan, RI10k, R30k, RI50k, and RI100k. Several, but not all, models were installed for multiple testing regions and will likely be installed later on in testing regions outside Japan. Testing models in multiple testing regions will lead to quicker assessments of the model performance. A brief description of each model is given by Nanjo *et al.* (2011). Figures 1–3 show maps of earthquake rates, λ (number / 3 months), for $M \geq 4$ in the “All Japan,” “Mainland,” and “Kanto” regions, respectively. A reference RANDOM forecast model was included into each of the tests applied to the three testing regions. We randomized forecast rates of earthquakes by constraining the sum of the forecast rates to be equal to the total number of observed earthquakes, ignoring the Gutenberg-Richter relation (Gutenberg and Richter, 1944). Therefore, by definition, this is not an informative model to forecast locations and magnitudes of earthquakes but serves as the lowest baseline for comparison with other

Fig. 1. Maps of the expected numbers of $M \geq 4$ earthquakes, λ , (color coded) in the "All Japan" testing region for the forecast period 1 November 2009 to 31 January 2010. In the upper left of each image, the model name is given. Circles show $M \geq 4$ earthquakes in the testing region during the same period. Size scales with magnitude.

Fig. 2. Same as Fig. 1 for the “Mainland” region.

Fig. 3. Same as Fig. 1 for the “Kanto” region.

Fig. 4. Observed number of earthquakes in non-overlapping 3-month intervals since 1990. (a) “All Japan”, (b) “Mainland”, and (c) “Kanto”. The solid line shows the mean number along with two-sided bounds estimated from the standard deviation (dashed line) and two-standard deviations (dotted line). A filled square and a gray square in each figure show the number of earthquakes in the forecast period we study here and in the next 3-month period, respectively.

more meaningful models.

4.3 Results

In the forecast period for the first test, 115, 15, and 14 $M \geq 4$ earthquakes have occurred in “All Japan” (Fig. 1), “Mainland” (Fig. 2), and “Kanto” (Fig. 3), respectively. To investigate whether or not these numbers are significantly low or high compared to the long term average, we compare them to the number of $M \geq 4$ earthquakes observed in each of the non-overlapping 3-month intervals since 1 January 1990. In the “All Japan” testing region, the average is 128.4 and the standard deviation is 54.9 (top panel of Fig. 4). The 115 observed earthquakes in the forecast period of 1 November 2009–31 January 2010 are indicated by a filled square. Visual comparison shows that the observed number of earthquakes is close to the average number of earthquakes in the same periods and within one standard deviation. Also included for comparison is the number of earthquakes in the next forecast period (1 February–30 April 2010, shown as a gray square), which we obtained from the JMA catalog with preliminary determined solutions for earthquake parameters (we confirmed that these are the same as events from the catalog with final solutions). The number of observed earthquakes is again within one standard deviation. The analysis for the testing regions “Mainland” and “Kanto” shows similar results in the middle and bottom panels of Fig. 4, respectively.

The summary results for “All Japan” are shown in Tables 1 and 2. Also included in Table 1 are the two columns: “Forecast number” to show the total number of forecast earthquakes and “Forec/Obs in number” to show

its ratio to the total number of observed earthquakes. The γ values show that no forecast is rejected (no bold value in the γ column). The Triple-S-Japan model predicts 149.27 earthquakes, resulting in a ratio of 1.30. This overprediction is indicated by its small δ_2 value. After we submitted the paper, the Testing Center and the modeler of the Triple-S-Japan (J. D. Zechar) realized that the code of this model for all testing regions was working incorrectly. However, they agreed that this paper be published. The modeler modified the code and submitted it to the Testing Center for future rounds. Based on the ζ values, the spatial distribution of the RI10k, RI30k, and RI50k forecasts is inconsistent with the observation. The κ values indicate that the forecast magnitude distributions are all consistent with the observation. All of the forecasts incorporate the well-known empirical scaling between frequency and magnitude, the Gutenberg-Richter relation (Gutenberg and Richter, 1944). As expected, the observation is consistent in the total number with the RANDOM forecast, but not in the magnitude and spatial distributions. Noticeably, the γ value indicates that this forecast cannot be rejected. As discussed later, the γ values are higher correlated with the total forecast number of earthquakes than with the spatial and magnitude forecasting. The models HIST-ETAS5pa, HIST-ETAS7pa, MARFS, MARFSTA, and RI100k pass all applied tests.

The T/W-test results shown in Table 2 provide a comparative evaluation of the forecasts. This test tests pairwise comparison: it compares each model with each of the others in turn. We confirmed that the results are completely symmetrical so that the results shown in Table 2

Table 1. The L-, N-, M-, and S-test results for the “All Japan” region. The columns “Forecast number” and “Forec/Obs in number” show the total number of forecast earthquakes and its ratio to the observed total number of earthquakes, respectively. Bold values indicate that the observed target catalog is inconsistent with the corresponding forecast. The values for the RANDOM forecasts are highlighted in light gray.

All Japan	L-test	N-test		M-test	S-test	Forecast number	Forec / Obs in number
Model	γ	δ_1	δ_2	κ	ζ		
HIST-ETAS5pa	0.901	0.945	0.065	0.950	0.054	132.69	1.15
HIST-ETAS7pa	0.965	0.945	0.066	0.928	0.471	132.61	1.15
MARFS	0.638	0.661	0.373	0.992	0.050	119.18	1.04
MARFSTA	0.540	0.624	0.412	0.985	0.054	118.08	1.03
RI10k	0.505	0.832	0.192	0.974	0.000	125.29	1.09
RI30k	0.678	0.832	0.192	0.969	0.000	125.29	1.09
RI50k	0.740	0.832	0.192	0.984	0.000	125.29	1.09
RI100k	0.892	0.832	0.192	0.979	0.483	125.29	1.09
Triple-S-Japan	1.000	0.998	0.002	0.974	0.711	149.27	1.30
RANDOM	0.459	0.512	0.525	0.000	0.000	115.00	1.00

Table 2. T/W-test results for the “All Japan” region. Numbers in the columns correspond to the numbered models in rows. Left arrows in gray cells indicate that the row model is more informative than the column model at 95% confidence level. No arrow in gray cells indicates not-significance.

All Japan	Model	2	3	4	5	6	7	8	9
1. HIST-ETAS7pa		←	←	←	←	←	←	←	←
2. MARFS			←	←	←	←	←	←	←
3. MARFSTA				←	←	←	←	←	←
4. HIST-ETAS5pa					←	←	←	←	←
5. RI10k						←	←	←	←
6. RI30k							←	←	←
7. RI100k								←	←
8. RI50k									←
9. Triple-S-Japan									

are single-sided. The result shows that the HIST-ETAS7pa model is significantly more informative than the other models. The second-most informative model is the MARFS model. The top four informative models, which are HIST-ETAS7pa, MARFS, MARFSTA, and HIST-ETAS5pa are consistent with the observations in all the consistency tests. The Triple-S-Japan model is less informative than the others because of the incorrect code as discussed above.

The summary results for the “Mainland” region are shown in Tables 3 and 4. The γ values indicate that no forecast fails the L-test. The EEPAS and PPE models predict 30.98 and 30.79 earthquakes, respectively. This is too many compared to the 15 observed earthquakes (see also their small δ_2 values). However, comparison with past seismicity (middle panel of Fig. 4) shows that these forecast numbers are within one standard deviation. Thus, these models do not forecast extremely high values. The Triple-S-Japan model again predicts too many earthquakes because of the aforementioned reason. Similar to the case of “All Japan,” no model is rejected in the M-test (in the κ column). Table 3 indicates that the MARFS, MARFSTA, and RI10k models are consistent with the observations in all tests. The T/W-test indicates that the former two forecasts (MARFS and MARFSTA) are more informative than all other forecasts.

The summary results for the “Kanto” region are shown in Tables 5 and 6. One feature of the results is that every model is rejected at least once in Table 5, but comparison

among the models (Table 6) clearly shows that the RI10k model is more informative than the others.

One surprising result is that the RANDOM forecasts in all three testing regions are not rejected in the L-test. This implies that the result obtained from the L-test is highly dependent on the number of events rather than on their spatial or magnitude distribution. Figure 5 shows direct comparisons between pairs of quantile scores: (a) δ_1 vs. γ , (b) δ_2 vs. γ , (c) δ_1 vs. δ_2 , (d) κ vs. γ , and (e) ζ vs. γ . Squares show the Japanese data: data from the “All Japan,” “Mainland,” and “Kanto” testing regions are colored in red, green, and blue, respectively. Also included for comparison are the results of the RELM experiment indicated by white symbols (Schorlemmer *et al.*, 2010b; Zechar *et al.*, 2010a): the mainshock class (circle); the mainshock+aftershock class (diamond); the mainshock.corrected class (upward-pointing triangle); and the mainshock+aftershock.corrected class (downward-pointing triangle). From Figs. 5(a) and 5(b), we see that the quantile scores δ_1 and δ_2 of the N-test correlate with γ of the L-test. Figure 5(c) showing a negative correlation between δ_1 and δ_2 , supports the results in Figs. 5(a, b). From these correlations, models that have overpredicted earthquakes as seen in small δ_1 and large δ_2 are likely not to be rejected in the L-test. In contrast, Figs. 5(d) and 5(e) show weak correlations of γ with κ of the M-test and ζ of the S-test. The RELM results are consistent with the Japanese results, supporting our observed correlations. Based on their heuristic discussion, Zechar *et al.* (2010a) thought of the L-test

Table 3. Same as in Table 1 for “Mainland”.

Mainland	L-test	N-test		M-test	S-test	Forecast number	Forec / Obs in number
Model	γ	δ_1	δ_2	κ	ζ		
EEPAS	0.999	0.999	0.001	0.583	0.024	30.98	2.07
MARFS	0.361	0.271	0.810	0.809	0.235	12.45	0.83
MARFSTA	0.374	0.297	0.787	0.825	0.236	12.73	0.85
PPE	0.999	0.999	0.001	0.635	0.686	30.79	2.05
RI10k	0.683	0.770	0.312	0.767	0.027	17.68	1.18
RI30k	0.641	0.770	0.312	0.767	0.000	17.68	1.18
RI50k	0.685	0.770	0.312	0.775	0.001	17.68	1.18
RI100k	0.788	0.770	0.312	0.766	0.012	17.68	1.18
Triple-S-Japan	0.999	0.999	0.003	0.967	0.149	29.45	1.96
RANDOM	0.426	0.535	0.568	0.000	0.000	15.00	1.00

Table 4. Same as in Table 2 for “Mainland”.

Mainland	Model	2	3	4	5	6	7	8	9
1. MARFS	←	←	←	←	←	←	←	←	←
2. MARFSTA		←	←	←	←	←	←	←	←
3. PPE			←	←	←	←	←	←	←
4. RI10k				←	←	←	←	←	←
5. RI30k					←	←	←	←	←
6. RI100k						←	←	←	←
7. EEPAS							←	←	←
8. Triple-S-Japan								←	←
9. RI50k									←

Table 5. Same as in Table 1 for “Kanto”.

Kanto	L-test	N-test		M-test	S-test	Forecast number	Forec / Obs in number
Model	γ	δ_1	δ_2	κ	ζ		
HIST-ETAS5pa	0.072	0.284	0.801	0.268	0.000	11.67	0.83
HIST-ETAS7pa	0.086	0.287	0.798	0.273	0.000	11.70	0.84
RI10k	0.050	0.023	0.989	0.376	0.934	7.57	0.54
RI30k	0.019	0.023	0.989	0.338	0.164	7.57	0.54
RI50k	0.019	0.023	0.989	0.322	0.005	7.57	0.54
RI100k	0.008	0.023	0.989	0.365	0.000	7.57	0.54
Triple-S-Japan	1.000	1.000	0.000	0.731	0.968	37.38	2.67
RANDOM	0.437	0.536	0.570	0.001	0.000	14.00	1.00

Table 6. Same as in Table 2 for “Kanto”.

Kanto	Model	2	3	4	5	6	7
1. RI10k	←	←	←	←	←	←	←
2. Triple-S-Japan		←	←	←	←	←	←
3. RI30k			←	←	←	←	←
4. RI50k				←	←	←	←
5. HIST-ETAS7pa					←	←	←
6. HIST-ETAS5pa						←	←
7. RI100k							←

as comprising the N-test, which tests the rate forecast; the M-test, which tests the magnitude forecast; and the S-test, which tests the spatial forecast. Our detailed analysis supports their thought, and indicates that the L-test is higher weighted on the N-test than on the S- and M-tests. The RANDOM forecasts are included in Fig. 5 and surrounded

by open squares. Their quantile scores are consistent with the observed parameter correlations.

5. Discussion and Conclusion

The main purpose of this article is to give a summary of the CSEP Japan Testing Center and first test results. We

Fig. 5. Correlations among parameters. (a) δ_1 vs. γ , (b) δ_2 vs. γ , (c) δ_1 vs. δ_2 , (d) κ vs. γ , and (e) ζ vs. γ . Squares show the Japanese data: “Mainland” (green); “All Japan” (red); and “Kanto” (blue). Note that the points from the RANDOM forecasts are also included in each plate and surrounded by an open square. Uncolored symbols show the RELM data (Schorlemmer *et al.*, 2010b; Zechar *et al.*, 2010a): the mainshock class (circle); the mainshock+aftershock class (diamond); the mainshock.corrected class (upward-pointing triangle); and the mainshock+aftershock.corrected class (downward-pointing triangle).

present result obtained from the first round of the 3-month class (the forecast period is 1 November 2009–31 January 2010), applied to the 3 testing regions “All Japan,” “Mainland,” and “Kanto” in the Japanese earthquake forecast testing experiment within the CSEP framework. An overview of the experiment is given by Nanjo *et al.* (2011).

Like in all other CSEP experiments, we perform rigorous prospective tests only: The forecasts of each model are tested against future seismicity only; no prior knowledge was available during the creation of the forecasts. Prospective testing is the only way to test the predictive power of forecast models without any conscious or unconscious bias.

As shown for “All Japan” and “Mainland,” informative models in the T/W-test (Tables 2 and 4) generated consistent forecasts with the observations in the L-, N-, M-, and S-tests (Tables 1 and 3). However, the consistency test results are not easily used to determine highest and lowest forecasting accuracy among participating models. The use of the T/W-test, which provides a comparative evaluation of the forecasts, can help us overcoming this situation. The results for the “Kanto” region in Tables 5 and 6 show a good example of it. Obtaining such kind of comparative results is one of the CSEP goals.

While the Japan Testing Center could explore the submitted codes to understand the different hypotheses at work (and their implications), it may be more efficient to communicate directly with the modelers, who ought to understand their models reasonably well and be able to explain any shortcomings they might demonstrate in this first experiment. Consistent with this motivation, this paper was successfully polished by modelers’ comments on an earlier-version manuscript that reported the test results with only

minimal interpretation. We stress that it should not be assumed that the results in Tables 1–6 reflect the relative value of the models for prospective testing. This is only the first trial and more trials need to be attempted for an understanding of their comprehensive and comparative value, which will be discussed in future publications (Yokoi *et al.*, in preparation).

Perhaps, the fact that all registered models are tested in the same environment under the same rules is more important. It shows the availability of multiple testable models that meet the scope of the current national program of earthquake prediction research, as described in the section “Introduction.” This also suggests the possibility to design mixtures of different models, which could potentially be more informative than any of the individual forecast models taken alone. Possible mixtures include weighted combinations of individual models, with weights to be assigned to each model according to test results. There have so far been few mixture models: rare examples include two models presented by Rhoades and Gerstenberger (2009). They found that the optimal mixture based on a California version of the EEPAS model and a Short-Term Earthquake Probability (STEP) forecasting model for forecasting $M \geq 5.0$ earthquakes in California gave an average probability gain of more than 2 compared to each of the individual models. However, designing a mix of more than two models still presents a challenge. As shown in this paper, the interpretation of test results is not always obvious. If weights could be assigned, for example in a Bayesian sense, then such mixed models would be crucial for creating forecast models with large enough probability gains to have societal impacts. This approach then again needs to be tested within

a CSEP-type framework, where the performance of a given forecasting model can be tested objectively in a verifiable way.

Acknowledgments. The authors would like to thank two reviewers (J. D. Zechar and K. Yamaoka) for helpful comments. We thank the following colleagues for their contribution to the forecast models discussed in this paper: Y. Ogata, D. Rhoades, C. Smyth, and J. D. Zechar. Discussions with R. Console, B. Enescu, G. Falcone, F. Hirose, Q. Huang, M. Imoto, T. Ishibe, T. Iwata, M. Kamogawa, A. M. Lombardi, K. Maeda, W. Marzocchi, S. Matsumura, M. Murru, T. Nagao, M. Okada, K. Shimazaki, S. Toda, S. Uyeda, K. Yamashina, and J. Zhuang were beneficial. Special thanks go to M. Liukis and S. Yokoi for help in running the CSEP Japan Testing Center and the Japan Meteorological Agency for the permission to use the earthquake catalog. We also thank the open source community for the Linux operating system and the many programs used to produce this paper. Maps were created using the Generic Mapping Tools (Wessel and Smith, 1998). This work was conducted under the auspices of the Special Project for Earthquake Disaster Mitigation in Tokyo Metropolitan Area of the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

References

- Field, E. D., Overview of the working group for the development of Regional Earthquake Likelihood Models (RELM), *Seismol. Res. Lett.*, **78**(1), 7–16, doi:10.1785/gssrl.78.1.7, 2007.
- Gerstenberger, M. C. and D. A. Rhoades, New Zealand Earthquake Forecast Testing Centre, *Pure Appl. Geophys.*, **167**(8–9), 877–892, doi:10.1007/s00024-010-0082-4, 2010.
- Gerstenberger, M. C., D. A. Rhoades, M. W. Stirling, R. Brownrigg, and A. Christophersen, Continued development of the New Zealand earthquake forecast testing centre, *GNS Science Consultancy Report 2009/182*, GNS Science, Lower Hutt, 2009.
- Gutenberg, B. and C. F. Richter, Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.*, **34**(4), 185–188, 1944.
- Hirata, N., Past, current and future of Japanese national program for earthquake prediction research, *Earth Planets Space*, **56**, xliii–l, 2004.
- Hirata, N., Japanese national research program for earthquake prediction, *J. Seismol. Soc. Jpn. (Zisin)*, **61**, S591–S601, 2009 (in Japanese with English abstract).
- Japan Meteorological Agency, Revision of JMA magnitude, *Newslett. Seismol. Soc. Jpn.*, **15**(3), 5–9, 2003 (in Japanese).
- Jordan, T. H., Earthquake predictability, brick by brick, *Seismol. Res. Lett.*, **77**(1), 3–6, doi:10.1785/gssrl.77.1.3, 2006.
- Marzocchi, W., D. Schorlemmer, and S. Wiemer, Preface, *Ann. Geophys.*, **53**(3), III–VIII, doi:10.4401/ag-4851, 2010.
- Nanjo, K. Z., Earthquake forecasts for the CSEP Japan experiment based on the RI algorithm, *Earth Planets Space*, **63**(3), 261–274, doi:10.5047/eps.2011.01.001, 2011.
- Nanjo, K. Z., T. Ishibe, H. Tsuruoka, D. Schorlemmer, Y. Ishigaki, and N. Hirata, Analysis of the completeness magnitude and seismic network coverage of Japan, *Bull. Seismol. Soc. Am.*, **100**(6), 3261–3268, doi:10.1785/0120100077, 2010a.
- Nanjo, K. Z., D. Schorlemmer, J. Woessner, S. Wiemer, and D. Giardini, Earthquake detection capability of the Swiss Seismic Network, *Geophys. J. Int.*, **181**(3), 1713–1724, doi:10.1111/j.1365-246X.2010.04593.x, 2010b.
- Nanjo, K. Z., H. Tsuruoka, N. Hirata, and T. H. Jordan, Overview of the first earthquake forecast testing experiment in Japan, *Earth Planets Space*, **63**(3), 159–169, doi:10.5047/eps.2010.10.003, 2011.
- Obara, K., K. Kasahara, S. Hori, and Y. Okada, A densely distributed high-sensitivity seismograph network in Japan: Hi-net by National Research Institute for Earth Science and Disaster Prevention, *Rev. Sci. Instrum.*, **76**, 021301, doi:10.1063/1.1854197, 2005.
- Ogata, Y., Significant improvements of the space-time ETAS model for forecasting of accurate baseline seismicity, *Earth Planets Space*, **63**(3), 217–229, doi:10.5047/eps.2010.09.001, 2011.
- Rhoades, D. A., Application of a long-range forecasting model to earthquakes in the Japan mainland testing region, *Earth Planets Space*, **63**(3), 197–206, doi:10.5047/eps.2010.08.002, 2011.
- Rhoades, D. A. and M. C. Gerstenberger, Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.*, **99**(2A), 636–646, doi:10.1785/0120080063, 2009.
- Rhoades, D. A., D. Schorlemmer, M. C. Gerstenberger, A. Christophersen, J. D. Zechar, and M. Imoto, Efficient testing of earthquake forecasting models, *Acta Geophys.*, **59**(4), 728–747, doi:10.2478/s11600-011-0013-5, 2011.
- Schorlemmer, D. and M. Gerstenberger, RELM testing center, *Seismol. Res. Lett.*, **78**(1), 30–36, doi:10.1785/gssrl.78.1.30, 2007.
- Schorlemmer, D. and J. Woessner, Probability of detecting an earthquake, *Bull. Seismol. Soc. Am.*, **98**(5), 2103–2117, doi:10.1785/0120070105, 2008.
- Schorlemmer, D., S. Wiemer, and M. Wyss, Variations in earthquake-size distribution across different stress regimes, *Nature*, **437**, 539–542, doi:10.1038/nature04094, 2005.
- Schorlemmer, D., M. Gerstenberger, S. Wiemer, and D. D. Jackson, Earthquake likelihood model testing, *Seismol. Res. Lett.*, **78**(1), 17–29, doi:10.1785/gssrl.78.1.17, 2007.
- Schorlemmer, D., N. Hirata, F. Euchner, Y. Ishigaki, and H. Tsuruoka, A Probabilistic Completeness Study in Japan, The 7th General Assembly of Asian Seismological Commission, Y3-214, 2008.
- Schorlemmer, D., A. Christophersen, A. Rovida, F. Mele, M. Stucchi, and W. Marzocchi, Setting up an earthquake forecast experiment in Italy, *Ann. Geophys.*, **53**(3), 1–9, doi:10.4401/ag-4844, 2010a.
- Schorlemmer, D., J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, T. H. Jordan, and the RELM Working Group, First results of the Regional Earthquake Likelihood Models experiment, *Pure Appl. Geophys.*, doi:10.1007/s00024-010-0081-5, 2010b.
- Smyth, C. and J. Mori, Statistical models for temporal variations of seismicity parameters to forecast seismicity rates in Japan, *Earth Planets Space*, **63**(3), 231–238, doi:10.5047/eps.2010.10.001, 2011.
- Tsuruoka, H., N. Hirata, D. Schorlemmer, F. Euchner, and T. H. Jordan, CSEP Earthquake Forecast Testing Center for Japan, *Eos Trans. AGU Fall Meet. Suppl.*, **89**(53), S33A-1935, 2008.
- Wessel, P. and W. H. F. Smith, New, improved version of Generic Mapping Tools released, *Eos Trans. AGU*, **79**(47), 579, 1998.
- Wiemer, S. and M. Wyss, Mapping spatial variability of the frequency-magnitude distribution of earthquakes, *Adv. Geophys.*, **45**, 259–302, doi:10.1016/S0065-2687(02)80007-3, 2002.
- Zechar, J. D. and T. H. Jordan, Simple smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.*, **53**(3), 99–105, doi:10.4401/ag-4845, 2010.
- Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades, Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.*, **100**(3), 1184–1195, doi:10.1785/0120090192, 2010a.
- Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan, The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurrency Computat.: Pract. Exper.*, **22**(12), 1836–1847, doi:10.1002/cpe.1519, 2010b.

H. Tsuruoka (e-mail: tsuru@eri.u-tokyo.ac.jp), N. Hirata, D. Schorlemmer, F. Euchner, K. Z. Nanjo, and T. H. Jordan