ISSN 8755-6839

SCIENCE OF **TSUNAMI HAZARDS**

Journal of Tsunami Society International

Volume 38 **Number 4** **2019**

BETA TESTING FOR INCREASED ACCURACY AND IMPROVED PERFORMANCE OF THE INDONESIAN TSUNAMI EARLY WARNING APPLICATION (Ina-TEWA)

Madlazim[1,2], Tjipto Prastowo[1,2]

[1]Physics Department, The State University of Surabaya, Surabaya 60231, Indonesia
[2]Center for Earth Science Studies, The State University of Surabaya, Surabaya 60231, Indonesia
Correspondence: madlazim@unesa.ac.id

ABSTRACT

Beta testing is a final stage for product examination following a series of product development. Regarding this, beta testing performed for an increased accuracy of Indonesian tsunami early warning application (Ina-TEWA) was applied to real conditions of Indonesian earthquakes with magnitudes of 6.0 or greater that occurred during a time period from January 1, 2018 to July 1, 2019. This stage was conducted in real-time for use of tsunami early warnings officially managed by Centre for Research and Development, The Indonesian Agency for Geophysics, Climatology and Meteorology (BMKG). It is frequently preceded by a series of preliminary tests known as alpha testing usually designed only for laboratory-scaled examination that is limited and offline. While implemented in real-time tsunami warnings, beta testing is considered as examination of acceptance given by and hence direct feedback from users. The launch for this tsunami early warning application is of great importance in the sense that it measures directly tsunami parameters from the earthquake source. The main aim of this stage is to identify possible errors, if any, and to make them the errors minimum for an increased accuracy in tsunami assessment to a maximum value possible. Based on beta testing applied to 35 events during the time period examined here, it was merely 7 errors occurred and found for tsunami false warnings. It follows that the Ina-TEWA prediction for accurate tsunami warnings reaches to a value of 80% or false warnings by this application remains relatively high of 20%. Therefore, tsunami early warning application for future use requires a further increase in terms of accuracy.

Keywords: *alpha testing, beta testing, tsunami warnings, Ina-TEWA*

1. INTRODUCTION

Accurate and quick release of tsunami early warnings following major tsunamigenic earthquakes generated by tectonic movement may prevent local communities and surrounding infrastructures from severe catastrophes (Satriano et al., 2011; Suppasri et al., 2015). Soon after tsunami wave generation, tsunami passage in the open ocean and its continuing arrival times at coastal regions both near and far from the epicenter need to be predicted accurately (Prastowo et al., 2018) using rapid assessment of observed waveforms recorded by instrument, such as ocean bottom electro-magnetometer deployed at the ocean floor, Deep-ocean Assessment and Reporting of Tsunami (DART) surface buoys offshore, or tide gauges onshore. Within this context, tsunami waveform analysis requires a reliable software or application, which is technically developed for this matter.

Software development towards a reliable application or programming package includes a number of consecutive stages, that is, analysis of requirement, design of development, integration and testing, and deployment and maintenance. With respect to these, this study focuses on beta testing as part of sequential procedures for an increased accuracy and hence improved performance of the Indonesian tsunami early warning system (Ina-TEWS, a former system of tsunami warnings before Ina-TEWA), which has been operated since 2008 by the Indonesian Agency for Geophysics, Climatology and Meteorology (BMKG). However, as addressed by Madlazim and Prastowo (2016), the Ina-TEWS performance remains relatively low in terms of accuracy thereby requiring further improvement.

Kocbek and Heričko (2013) argued for the quality of a system that is driven by software-assisted performance. The performance is determined by a number of factors, such as accuracy, time-lapse, completeness, consistency, maintenance, security, safety, reliability, and usability. During the stages of system improvement, there are opportunities to find possible errors in various stages of product examination and development. Testing is thus generally applied for a specific purpose as a main tool to ensure a software of high quality, providing specific methods and corresponding techniques for error detection in such a system (Budnik, 2012; Kocbek and Heričko, 2013). However, software bugs are always present because the complexity of a developed software in general cannot be fully solved. It follows that design limitation can never be completely avoided, in particular when dealing with complex software-supported systems (Pan, 1999; Mohd and Shahbodin, 2015).

For improvement of performance, software testing in any stage of product development is carried out normally for the following main objectives: (1) quality assurance; (2) verification and validation; and (3) reliability estimation, where each of these refers to the suitability of specified product design requirement. When these are achieved, the high quality of product is attained in some circumstances. Debugging, a common way of software testing, is usually carried out to search for weaknesses of a program hence assuring the program being able to run well during performance. This process is of primary importance as it is impossible to produce an adequately complex program left uncorrected in its first-time use. According to Madlazim et al. (2015), another important purpose of software testing is verification and validation. In the process of fulfilling this purpose, software testing is a measure of its applicability hence reliability. Software reliability is therefore directly connected to the number of tests completed. Within the uncertainty during test measurements, software testing can therefore be used to effectively collect a set of data failures or unsuccessful performance for estimates of product reliability in future use (Pan, 1999; Madlazim et al., 2015).

As previously mentioned, software testing is complex, containing a series of tests. A preliminary test called alpha testing is performed by prospective users or independent programmers involved in the product development division. The primary aims of this test are to detect any error in early stages of development and to bring the corrected syntax embedded in a software to a further test for detailed examination, namely beta testing (Pradhan, 2012). In the present study, beta testing of the Ina-TEWA is reported. The primary purpose is to identify how many tests required for optimum performance to reveal a number of design defects in this tsunami early warning application and to reduce them all to a minimum since the application performance is measured in terms of its accuracy. We present here beta testing for 35 events during a time period from January 1, 2018 to July 1, 2019 in the country, recorded by a network of seismic stations managed by BMKG and then analysis all of the results. Discussions on these results are given before concluding remarks at the end of this work.

2. INDONESIAN TSUNAMI EARLY WARNING APPLICATION (Ina-TEWA)

Tsunami monitoring application in Indonesian territories, formally abbreviated as the Ina-TEWA, has currently used main earthquake-tsunami discriminants, such as rupture duration T_{dur}, the duration exceed 65 seconds $T_{65\mathrm{ex}}$ and the dominant period T_{d} of the first arrivals of P-waves following a series of pioneering work by Lomax and Michelini (2009, 2011, 2012). The primary difference between the Ina-TEWA and these studies is the use of a network of local stations for the Ina-TEWA, rather than teleseismic observations, to record seismic activities generated by earthquakes of varying magnitudes. In the early stages before the Ina-TEWA was applied, the Ina-TEWS extracted seismic signals from a local network nearby an event and transformed it into a digital set of source parameters ready for analysis and assessment, including earthquake magnitude, origin time, epicenter and hypocenter. However, as claimed by Madlazim and Prastowo (2016), most Indonesian earthquakes with epicenters in the ocean with depths shallower than 70 km and magnitudes greater than 7.0 during 2007-2010, were falsely announced as tsunamigenic events. This requires evaluation of earthquake parameters used in the Ina-TEWS as a result of inaccuracy in tsunami assessment. This inaccuracy led to false warnings either tsunami alerts were issued as rapid responses to earthquakes of adequately large magnitudes or likewise occurrences with relatively small magnitudes were left to generate tsunami hence a series of destruction on its way to beaches and lands with no warnings. In regard to all of these false warnings, quick action after the main shocks is not the only important issue since accuracy in providing information is also significant to earthquake-tsunami assessment.

To reduce the level of inaccuracy in false warnings formerly issued by BMKG for tsunami events in Indonesian regions, Madlazim et al. (2015) have then utilized numerical codes called *Joko Tingkir* for a better prediction of tsunami wave generation that may occur following ocean bottom ruptures by future events. These codes have incorporated the importance of geometry of a fault in parameterizing tsunami potential into play, which includes the length and width of a rupture. Since the length of a rupture is technically difficult to measure, T_{dur} that is proportional to the length of a rupture is then used for a discriminant. Together with this parameter, $T_{65\mathrm{ex}}$ and T_{d} (determined from seismograms), are likely being used for a set of earthquake-tsunami discriminants for effective tsunami warnings. In addition to tsunami importance reported by Lomax and Michelini (2011), for events in Indonesian regions, Madlazim et al. (2015) argued that the discriminants are as follows, $T_{\mathrm{dur}} \geq 65$ s, $T_{\mathrm{d}} \geq 10$ s, $T_{65\mathrm{ex}} \geq 1$, $T_{\mathrm{dur}} \times T_{\mathrm{d}} \geq 650$ s^2 and $T_{\mathrm{d}} \times T_{65\mathrm{ex}} \geq 10$ s. If at least three of these inequalities are satisfied then a tsunami wave is possible and hence to issue tsunami alert.

3. METHOD FOR BETA TESTING

While being integrated into the existing monitoring system to support the Ina-TEWS performance with a direct supervision from BMKG, a programming package named *Joko Tingkir* application was validated using tsunami importance (Madlazim et al. 2015) and then tested using alpha testing during a period of 2014-2015 for early detection of possible errors. The results showed that the application has fulfilled requirement in an adequate level and confirmed that further improvement of performance for an increased accuracy hence perfectness is required. This is an essence of the current study, where beta testing of the Ina-TEWA is presented.

In the process of improvement in terms of accuracy, Madlazim et al. (2018) developed a method called Mfilter used to reduce seismic signals from unexpected sources of dominant noises. Although this method was not directly specified to increase the accuracy, it helped the Ina-TEWS performance better in that automatic computation of *P*-waves run faster, making the calculation process efficient (see for further use of automation in Kirubakaran and Karthikeyani, 2013). Therefore, beta testing was performed during the progress of this study whereas such a direct method of computation was also implemented in the automated system. For completeness, beta testing utilized a total of 35 earthquake events throughout Indonesian territories and its surrounding regions for the last two-years occurrences having magnitudes of 6.0 or greater and sources of 10 to 600 km deep beneath the surface, as depicted in Figure 1 below, where the majority of the events were located in the Indonesian eastern provinces and some of these occurred in Philippine and Papua New Guinea (PNG). Real-time measurements of the 5 earthquake-tsunami discriminants were carried out during the test using methodology detailed in previous work of Madlazim et al. (2015).

Before going further into the next section, here we provide the saline points of what is referred to as tsunami false warning. By definition, a false warning used in this study means that the Ina-TEWA suggests TSUNAMI POTENTIAL for an event of a particular magnitude, either hypocentered below the ground or the sea surface at a great depth, from which a tsunami alert may then be publicly issued by the relevant Indonesian authorities although field observations confirm no tsunami generated or the application says NO TSUNAMI POTENTIAL even though monitoring instruments offshore and onshore confirm tsunami generation. It is relatively easy to understand that the two possible ways of false warnings have to be avoided for putting people living and all properties nearby local beaches in danger. Even in some cases, places far away from the source remains vulnerable to tsunami hazard as the wave may travel a great distance across the ocean (Prastowo et al., 2018).

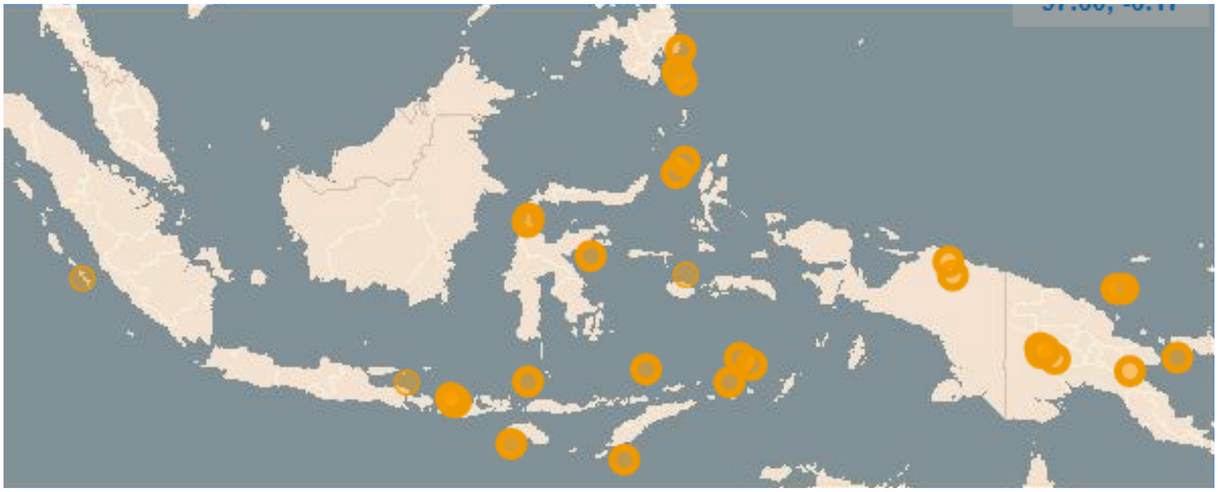

Figure 1. Distribution of epicenter of 35 earthquakes with magnitudes 6.0 or greater (marked orange circles) and varying depths from shallow to deep sources examined in this study during a time period of January 1, 2018 till July 1, 2019.

4. RESULTS AND DISCUSSIONS

The Indonesian tsunami early warning application project has recently been extended into three versions that are managed by BMKG to date. The first version of the application was published in June 2013, followed by the second implementation in March 2016, and the third in November 2017. For each, a comprehensive computer testing was performed, as required. When the application is publicly used, there is no a turning point. Hence, this is why the process of internal-driven testing is of importance. For this reason, a group of people called an internal team responsible for beta testing (Buskey, 2005) has examined daily records on the use of the application, subsequently reviewing them all in detail. All errors, shortcomings, inconsistencies and feature requests were recorded by a Flyspray system, freely accessible at http://flyspray.org/ popularly known as a web-based bug tracking system, written in the form of a hypertext preprocessor for software development. The system allowed users to record few characteristics of the application, including a reported version, a place of testing, a severity of test results, a priority of usability problems, and examiner names.

The group contains researchers utilizing this system for beta testing routinely once within a 5-days period of time. An external group named the Centre for Research and Development BMKG as the official authority responsible for beta testing obtains a weekly report. This group has examined all detected defects due to any imperfect design of the application. Before product release or launch for implementation in a real condition, it is required for the two groups to report all the problems found in a manner different from ways of guidelines and reporting engineering case studies recommended by Runeson and Höst (2008).

As mentioned, an internal team for beta testing plays a key role in the software development. Examiners within the team carry out a number of automatic computer-based tests as well as manuals on the devices and for earthquakes occurring throughout Indonesian territories and its surroundings. The primary purpose of this step is to identify design defects as many as possible to ensure tsunami early warning application working with high performance hence high accuracy, as expected. In order to make it real, a number of people within the division of human resources capable of doing the job on demand are needed. These people have examined the application in details within a framework of previous studies (Nielsen and Landauer, 1993; Nielsen, 2000; Jiang et al., 2017), who suggested that the best result of beta testing for optimum performance is obtained from only 5 internal users.

Adopting the formula for the number of usability problems found in a particular test (Nielsen and Landauer, 1993) and applying it in the context of tsunami assessment, the number of false warnings X given by the application under examination in beta testing with n users is calculated from

$$X = N\left\{1 - (1 - L)^{n}\right\} \qquad (1)$$

where N is the total number of the problems detected in the software design and L is the proportion of design defects found when a single user is on work. A typical value of L is of 31% (Nielsen, 2000), on average, for a large number of projects under consideration.

Using the above value for L, the following results are then presented and subsequently discussed. The surprising result from Eq. (1) is that zero users with $n = 0$ give zero false warnings with $X = 0$, as expected when there is no implementation. When the data were collected from a single test user with $n = 1$, false warning rapidly increased to approximately one third of the total problems detected in the design. The large difference between a case of zero users and a few was deliberately shocking. When the second user was then in action (with $n = 2$), we found the same things being performed as the first single test user did, making an overlap in the learning process between the first two-users. However, there will always be something different from the two. For example, if the second user does a thing then the first user may not do it. Within this context, the second user provides new insight but it is not as much a small amount of new data, relatively compared with the previous two users. Keep it in mind, more and more users involved, then less and less lessons learned because the same things occur as the first user does. The third user (with $n = 3$) would then do many things compared with the first or the second user and even things observed twice, and again, would generate repeatedly. It is not necessary to keep the same things observed multiple times, and there exists enough reasons for being back to the drawing board and redesign the application to eliminate the usability problems. In short, by adding more users fewer false warnings are found because the same situation is repeated. After the fifth user, it is a time waste to observe the same problems or common findings repeatedly but not providing better insight much (Nielsen and Landauer 1993; Nielsen 2000). In other words, optimum performance of the test is achieved at a value of $n = 5$ (meaning that 5 users are involved) permitting us to set up the equal number of examiners for each group of internal and external origin during the test. Project on beta testing for the Ina-TEWA was carried out following such descriptions, where each group of examiners or testers was filled with 5 examiners or testers for the best results of optimum performance. Before testing was performed, the examiners or testers were knowledgeable by relevant instructions about which function of the application required to be tested. Each examiner or tester performed a number of tests, once every 5 days, where the Flyspray is the main logging tool. This device recorded a single design defect detected by an examiner or a tester during the test.

To facilitate data analysis, we consider the records collected in a week before the latest version (the third one) of the Ina-TEWA was launched. False warnings were classified into whether they were related to a level of accuracy in the application or a technical problem. All examiners or testers found 7 of 35 earthquakes examined in this study were falsely issued (5 problems of accuracy and 2 cases of inconsistencies). To validate this, we focus on the first 5 examiners or testers from the internal group and present the results in Table 1 below.

In consecutive, tester A found 7 errors in tsunami alert with no new false warnings reported. The same situation for tester B with zero new false reports issued, followed by tester D and E having the same experiences. Different from the others, tester C observed 6 tsunami alerts mistaken and a new error was recorded. We then move on the second 5 examiners or testers from the external group and present the results in a different panel below. A clearly similar performance is shown by the two groups for the same total number of events examined in this study, where the two have found 7 cases of falsely issued warning. It follows that 7 of 35 cases considered is categorized as false in tsunami alert reported. There remains challenging for future use of beta testing whether optimum performance using 5 testers for each examiner, as is the case in the present study, being efficient and consistent with more cases of earthquakes (Jiang et al. 2017).

Table 1. False warnings reported by 5 testers from internal (FGHIJ) and external (ABCDE) groups of researchers responsible for beta testing using methodology of optimum performance.

Tester	False Warning	New False Warning
A	7	0
B	7	0
C	6	1
D	7	0
E	7	0

Tester	False Warning	New False Warning
F	7	0
G	7	0
H	6	1
I	7	0
J	6	1

The occurrence of 20% false warnings found using the 2017 Ina-TEWA version remains relatively high. This is due to only the dominant period T_d of the first P-wave arrivals as the main parameter for tsunami detection being filtered using techniques described in Madlazim et al. (2018). The other two crucial parameters, T_{dur} and T_{65ex}, have not yet been directly calculated from filtered seismic signals, making them remains consisting of unexpected noises and thereby reducing the level of accuracy.

5. CONCLUSIONS

Beta testing for improved performance of the Ina-TEWA in terms of accuracy has been carried out using a total of 35 relatively large earthquakes of varying source depths in Indonesian and its nearby regions, particularly in the Indonesian eastern provinces where seismic activities have been found to increase lately. These events are all considered tsunamigenic considering their scales in magnitude. After a series of stages during the test, we analyze all the results and conclude that this test is proved to increase the level of accuracy and hence to improve the Ina-TEWA performance by 80% since only 7 tsunami warnings are falsely issued. However, further refinement of the existing tsunami warning application remains challenging for future work. In the light of perfectness within the uncertainty or possible errors permitted in the process of computation, upcoming research on this subject may then also install M-filter used for the other two main tsunami parameters (T_{dur} and $T_{65\mathrm{ex}}$).

Acknowledgements

The authors would sincerely like to thank anonymous reviewer(s) for their best suggestions and invaluable comments upon this manuscript to be appropriate for publication in STH. Great thanks go to the Center for Research and Development BMKG, Indonesia, the authority with which all the data used are obtained and from which permission to run the application is given for this study.

References

Budnik, C. (2012), Software testing, software quality and trust in softwarebased systems, *Proc. 36th Annual IEEE Computer Software and Applications Conference* (COMPSAC), 253, Izmir, Turkey: IEEE, doi:10.1109/COMPSAC.2012.107.

Buskey, C. D. (2005), A software metrics based approach to enterprise software beta testing design, *doctoral dissertation*, 1-268, unpublished work, New York, US: Pace University Press, ISBN: 0-542-34378-9.

Jiang, Z., K. P. Scheibe, S. Nilakanta and X. Qu (2017), The economics of public beta testing, *Deci. Sci*. doi:10.1111/deci.12221.

Kirubakaran, B. and V. Karthikeyani (2013), Mobile application testing: challenges and solution approach through automation, *Proc. International Conference on Pattern Recognition*, *Informatics and Mobile Engineering* (PRIME), 79-84, Periyar University, Tamilnadu, India: IEEE, doi:10.1109/ICPRIME.2013.6496451.

Kocbek, M. and M. Heričko (2013), Beta testing of a mobile application: a case study, *the Second Workshop on Software Quality Analysis, Monitoring, Improvement, and Applications* (SQAMIA), Novi Sad: Department of Mathematics and Informatics, University of Novi Sad, Novi Sad, Serbia.

Lomax, A. and A. Michelini (2009), Tsunami early warning using earthquake rupture duration, *Geophys. Res. Lett.*, 36, L09306, 1-4, doi:10.1029/2009GL037223.

Lomax, A. and A. Michelini (2011), Tsunami early warning using earthquake rupture duration and *P-wave* dominant period: the importance of length and depth of faulting. *Geophys. J. Int.*, 185(1), 283-291, doi:10.1111/j.1365-246X.2010.04916.x.

Lomax, A. and A. Michelini (2012), Tsunami early warning within five minutes. *Pure Appl. Geophys*., doi:10.1007/s00024-012-0512-6.

Madlazim, T. Prastowo and T. Hardy (2015), Validation of Joko Tingkir software using tsunami importance. *Sci. Tsu. Haz*., 34(3), 189-198.

Madlazim and T. Prastowo (2016), Evaluation of earthquake parameters used in the Indonesian Tsunami Early Warning System, *Earthq. Sci*., 29(1), 27-33, doi:10.1007/s11589-016-0143-6.

Madlazim, T. Prastowo, S. Rohadi and T. Hardy (2018), Filter-M application for automatic computation of P wave dominant periods for tsunami early warning. *Sci. Tsu. Haz*., 37(1), 26-32.

Mohd, CKNCK and F. Shahbodin (2015), Personalized learning environment: alpha testing, beta testing and user acceptance test, *Proc.Soc. Behavioral Sci*., 195, 837-843, doi:10.1016/j.sbspro. 2015.06.319.

Nielsen, J. (2000), Why you only need to test with 5 users. retrieved on September 2, 2013 from http://www.nngroup.com/articles/whyyouonlyneedtotestwith5users/.

Nielsen, J. and T. K. Landauer (1993), A mathematical model of the finding of usability problems. *Proc. ACM INTERCHI'93 Conference on Human Factors in Computing Systems*, 206-213, Amsterdam, The Netherlands: CM DL, doi:10.1145/169059.169166.

Pan, J. (1999), Software testing, retrieved on September 2, 2013 from http://www.ece.cmu.edu/~koopman/des_s99/sw_testing/

Pradhan, T. (2012), All types of software testing, retrieved on September 2, 2013 from http://www.softwaretestingsoftware.com/all typesofsoftwaretesting/

Prastowo, T., L. Cholifah and Madlazim (2018), Analysis of travel time delay for large tsunamis across the Pacific and Indian Oceans, *Sci. Tsu. Haz*., 37(4), 195-212.

Runeson, P. and M. Höst (2008), Guidelines for conducting and reporting case study research in software engineering. *Empir. Software Eng*., 14(2), 131-164, doi:10.1007/s10664-008-9102-8.

Satriano, C., Y. M. Wu, A. Zollo and H. Kanamori (2011), Earthquake early warning: concepts, methods, and physical grounds, *Soil Dyn. Earthq. Eng*., 31(2), 106-118, doi:10.1016/j.soildyn. 2010.07.007.

Suppasri, A., K. Goto, A. Muhari, P. Ranasinghe, M. Riyaz, M. Affan, E. Mas, M. Yasuda and F. Imamura (2015), A decade after the 2004 Indian Ocean tsunami: the progress in disaster preparedness and future challenges in Indonesia, Sri Lanka, Thailand, the Maldives, *Pure Appl. Geophys*., 172, 3313-3341, doi:10.1007/s00024-015-1134-6.