Bulletin of the Seismological Society of America, Vol. 93, No. 1, pp. 519–525, February 2003

Short Notes

Evaluating the Fit of Alternative Hypocenters to Arrival Times

by Raymond J. Willemann

Abstract Statistical methods for comparing travel time residuals are reviewed and the importance of using a statistic that is both robust and a measure of dispersion is demonstrated. Results from relocations based on travel times from the Jeffreys-Bullen (JB) tables and from a tomographic (3D) model are compared using the variance of the Winsorized samples, which is a robust measure of dispersion. The improvement of the residuals is significant at a high level of confidence globally and in many individual regions. The same events are relocated with travel times of the modern one-dimensional model ak135. The incremental improvements from JB to ak135 residuals and from ak135 to 3D residuals are each very small. Each of the small incremental improvements is statistically significant when all earthquakes from around the world are taken as a single sample, but regional samples of even hundreds of earthquakes rarely show that either of the incremental improvements is significant.

Introduction

Accurate hypocenters are required for a broad range of seismological purposes, ranging from hazard analysis to tomography. Hypocenters reported in catalogues are taken as sufficiently accurate for some applications, and even when requirements for more accurate hypocenters are recognized, relocations are often based on a linearization of the problem in the vicinity of catalogue hypocenters. Thurber and Engdahl (2000) summarize reasons why many catalogues still fall short of the desired accuracy and the efforts underway to improve accuracy using new algorithms and travel time models.

When a purportedly more accurate set of hypocenters is produced, judging the improvement is often problematic. Where *ground truth* locations are known, of course, a newly computed epicenter or depth that is closer to ground truth is superior (Sultanov *et al.*, 1999; Ritzwoller *et al.* 2000). But ground truth locations are often available only for events in some limited geographic or depth range, or from a time before installation of seismic stations now in use. Thus, for the great majority of events with no known ground truth location, some alternative evaluation would be useful.

It may seem self-evident that more accurate hypocenters should fit arrival times better (Engdahl *et al.*, 1998), but this simple statement is subject to numerous caveats. The major reason is that arrival time errors are variable and difficult to quantify (Thurber *et al.*, 2000). My objective is to demonstrate that carefully chosen statistical methods can be used to compare how well differently computed hypocenters fit arrival times despite the presence of outliers. The outliers that may be troublesome in such comparisons are those with residuals that have a non-Gaussian distribution but that may be as small as differences between actual travel times and travel times computed from a model. Recognition that arrival time errors are non-Gaussian extends back to the earliest attempt to use them to locate earthquakes and infer earth structure (Jeffreys, 1932). The errors arise partly from uncertainty in measuring times in seismic records, of course, but non-Gaussian errors have a wide variety of causes (Pavlis, 1992). Logical blunders such as mistaking prearrival noise for an arrival, misidentifying a phase type, and associating an arrival with the wrong event are ubiquitous, are not corrected by simply using an improved travel time model, and dominate some measures of misfit.

The statistical methods that I use fall within the classes of statistical estimators that have been tested in attempts to compute hypocenters that are more accurate or for which true uncertainties can be computed more confidently (Anderson, 1982). But in this article my objective is to judge how well two sets of hypocenters fit arrival times, by whatever means they were computed. For specific examples, I will use results from Chen and Willemann (2001), who recomputed hypocenters from *P* arrival times reported in the *Bulletin of the International Seismological Centre* (International Seismological Centre [ISC], 2000), hereinafter the *Bulletin*. In one of their tests, Chen and Willemann (2001) used the tomographic model of Kárason and van der Hilst (1999) to recompute hypocenters for approximately 3800 events reported in the *Bulletin* for 1998 January with *P* arrival times from a sufficiently good station distribution. Ground truth locations were not available for these events; the primary purpose of this particular test was to investigate the distribution of displacements from the locations in the

Bulletin based on the JB travel times (Jeffreys and Bullen, 1940). The new hypocenters and arrival time residuals are available at www.isc.ac.uk/~qfchen/.

Measures of Scale and Dispersion

Chen and Willemann (2001) report the root mean square (RMS) of residuals, r, for each recomputed hypocenter. Just as in the *Bulletin*, the reported rms value is a weighted mean,

$$s_U = \sqrt{\frac{n \quad \Sigma w r^2}{n - d \ \Sigma w}} \tag{1}$$

where n is the number of arrival times used to compute the hypocenter and d is the number of degrees of freedom in the solution (3 if depth is fixed, 4 otherwise). In the method of uniform reduction (Jeffreys, 1932), the weights, w, are computed as

$$w = \frac{1}{1 + \mu \exp(r^2/2\sigma^2)} \tag{2}$$

where, in turn, σ is an estimate of the standard deviation of arrival time residuals, and μ is a dimensionless parameter that controls how steeply w decreases for larger residuals. For both the *Bulletin* and Chen and Willemann (2001) $\mu = 0.05$ and, following Buland (1986), $\sigma = 1.145$ sec. The weights are recomputed at each iteration, which introduces additional nonlinearity. Use of a *redescending* weighting function such as Equation (2) may create local minima further to any that already exist in an inverse problem (Huber, 1996). Nevertheless, uniform reduction is a powerful technique for computing hypocenters from arrival times, which include an initially unknown fraction outliers and also preserve the desirable property of χ^2 distributed hypocentral errors (Buland, 1986).

A redescending weighting function also has the undesirable result of producing a measure of scale that is not a measure of dispersion (Wilcox, 1997). That is, s_U can increase as a result of reducing some fraction of the residuals. Suppose, for example, that large corrections are introduced for a few stations and improve the situation, so that a small proportion of the outliers are replaced by residuals with the same Gaussian distribution as the majority of the residuals. Numerical simulations easily confirm that this correction is almost as likely to increase s_U as decrease it. The increase can occur because the changed residuals that were outliers were assigned effectively 0 weights and made no appreciable contribution to s_U. But when the residuals are reduced to the distribution of the majority, they are equally likely to be larger or smaller than the other residuals that already contribute to s_U.

Both in the *Bulletin* and in the results from Chen and Willemann (2001), the mode of the distribution of s_U values for these events is close to the assumed value of σ (Fig. 1, left). Values of s_U much larger than σ are rare because residuals larger than σ are assigned smaller weights and will not contribute much to s_U unless there are many more large residuals than small residuals. Thus, the location of this mode is an artifact of the processing parameters. For example, if Chen and Willemann (2001) had supposed that residuals would necessarily be smaller with a better travel time model and so used a smaller value of σ, then the mean value of s_U would be cut regardless of whether fit to arrival times actually improved.

The artifacts inherent in s_U, as well as its failure to be a measure of dispersion, make it a poor choice for comparing how well two hypocenters fit the data, despite its undoubted utility in computing the hypocenters in the first place. We might consider comparing the fits using the unweighted rms residuals,

$$s_1 = \sqrt{\frac{\Sigma r^2}{n - d}}. \tag{3}$$

But s_1 exceeds 1 sec for most of Chen and Willemann's (2001) January 1998 hypocenters. The problem with s_1 is that its dependence on any one residual is unbounded, that is, it is not robust, so an outlier can predominate its value.

To compare how well two hypocenters fit arrival times we require a statistic that avoids both the flaws of s_U and the flaws of s_1, that is, a robust measure of dispersion. One possibility that we might consider is a quantile, for example, the 68th percentile is a robust estimator of the standard deviation of the population of residuals. Under many circumstances, quantiles are an effective measure of misfit to use in inverting for an optimum solution, but they may be less appropriate for comparing two optimum solutions computed using different models. For example, if the outliers are caused by logical blunders then they are as likely to increase as decrease when a new travel time model is introduced. If such outliers are sufficiently numerous (e.g., more than 32% of the data if the 68th percentile is used), then the expected change of a quantile of the residuals would be 0 even for a genuinely better fit to the data that are not outliers.

Other robust measures of dispersion are not estimators of the population standard deviation at all, but they avoid the flaw of insensitivity to improving "good" residuals, that is, those residuals that are already better than a quantile. Two well-studied robust measures of dispersion are the variance of trimmed samples and the variance of Winsorized samples (Wilcox, 1997). A trimmed sample is simply the observed sample with some fixed fraction of the smallest and largest values removed. A Winsorized sample is the observed sample with a fixed fraction of the smallest and largest values replaced by values at the bounding quantiles (Table 1). The fixed fraction trimmed or Winsorized from each sample should exceed the fraction of values that might be outliers in any of the samples, so small trims are usually avoided; trimming or Winsorizing the smallest 20% and largest 20% of values in each sample is typical.

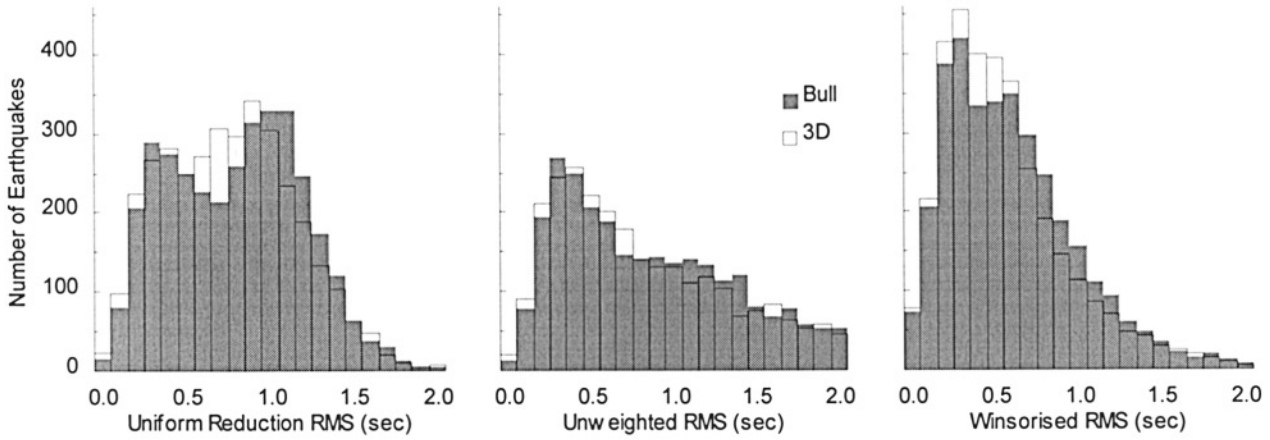

Figure 1. Histograms of rms residual reported for hypocenters of 1998 January events as reported originally in the *Bulletin* and by Chen and Willemann (2001). Each histogram includes one datum for each event. *Left*: s_u is computed using weights from Jeffreys' method of uniform reduction (Equations 1 and 2). *Center*: s_1 is computed without weights (Equation 3). *Right*: s_W is computed without weights but from the Winsorized residuals (Table 1). Some events have residuals in the *Bulletin* with rms values >2 sec, which is beyond the domain in this figure; six events in the case of Uniform Reduction rms, 826 events in the case of Unweighted rms, and 64 events in the case of Winsorized rms.

Table 1
Example of Winsorising and Trimming

	Full	Winsorized	Trimmed
	0.088	0.260	
	0.169	0.260	
	0.260	0.260	0.260
	0.405	0.405	0.405
	0.459	0.459	0.459
	0.610	0.610	0.610
	0.824	0.824	0.824
	0.841	0.841	0.841
	0.906	0.841	
	0.944	0.841	
Mean	0.551	0.560	0.567
rms	0.303	0.248	0.186

A sorted sample of ten pseudo-random values in [0, 1], the 20% Winsorized version of the full sample, and the 20% trimmed version of the full sample. *Mean* is the mean of the values in each sample; the mean of the Winsorized and trimmed versions of a sample are robust but non-maximum likelihood estimates of the population mean. *RMS* is the square root of the mean of the squared departure of each sample from its own sample mean. RMS of the Winsorized and trimmed samples are robust measures of the population dispersion, but neither of them is an estimate of the population standard deviation.

The square root of the variance of the 20% Winsorized sample, hereinafter s_W, is a true measure of dispersion, that is it cannot increase as a result of decreasing any residual or set of residuals. The salient feature that makes s_W a measure of dispersion is that the fraction Winsorized is fixed. Winsorizing only empirically identified outliers would nearly recover the method of uniform reduction but with an abrupt transition of the weight from 1 to 0. With a fixed fraction of each sample Winsorized, the worst possibility is that s_W might remain unchanged despite improvement. This failure to show improvement would occur if the only difference was that a few outliers were corrected, but not enough to bring them within the Winsorization quantile. For Chen and Willemann's (2001) January 1998 events, the mode of the distribution of s_W is unchanged (Fig. 1, right). That is, the new travel time model has not much improved the fit of data where residuals were already on the order of 0.3 sec. Among events with residuals originally on the order of 0.6 to 1.2 sec, however, the distribution of s_W shows that relocation with a better travel time model often improved the residuals by several tenths of seconds. This pattern is quite reasonable since measurement errors probably limit the extent to which very small misfits can be reduced, while larger misfits can occur for earthquakes and station distributions where the original travel time model can be significantly improved.

Significance of Differences

Unless we adopt a population distribution for residuals, it is difficult to state an uncertainty of s_W, or for an individual earthquake, a statistical significance of the change in s_W from the *Bulletin* to other results. Nevertheless, we can infer statistical significance from differences between s_W for a set of earthquakes. Specifically, we can measure of the likelihood that there is any difference at all between the two sets using the Kolmogorov–Smirnov statistic,

$$J_s = \max\left|F_s^{3D} - F_s^{Bull}\right| \quad (4)$$

where F_s is the sample cumulative distribution function of s_W. The Kolmogorov–Smirnov statistic is evaluated by defining the confidence that two samples differ as

$$C_s = 1 - P(J > J_s|n), \quad (5)$$

where the probability P is computed assuming that the two samples of n values were drawn independently from the same population. The reason for using the Kolmogorov–Smirnov test is that it is distribution-free, that is it is valid for any continuous population distribution (Hollander and Wolfe, 1999). For all of the 3484 events of January 1998 relocated by Chen and Willemann (2001) using arrival times at five or more stations J_s is 0.07. With such a large n this confirms that in aggregate $s_W{}^{3D}$ and $s_W{}^{Bull}$ differ with more than 99.9% confidence.

While the regional confidence levels of differences span the range from 0 to 1, in 12 of 48 regions the Kolmogorov–Smirnov test indicates a difference between $s_W{}^{3D}$ and $s_W{}^{Bull}$ at the 90% confidence level.

In regions of low seismicity, the number of earthquakes in 1 month is often insufficient to demonstrate statistically significantly improvement of the residuals. Where stations are sparse there is significant improvement even in some regions with small lateral heterogeneity, such as the Galapagos region and the Arctic. But in regions where stations are dense, improvement is not statistically significant even if earthquakes are numerous and lateral heterogeneity is large, such as Japan. One factor that might limit improvement in well-monitored regions is relatively small errors before relocation. Also, Chen and Willemann used the globally averaged model ak135 for travel times of local and regional phases such as *Pg* and *Pn*, which more strongly influence locations where stations are denser.

Evaluation of Individual Events

Even in regions where improvement is most significant, relocation with a new travel time model makes s_W worse for a few events. We cannot say directly whether or not dispersion is significantly greater because we have not estimated the uncertainty of s_W. But we can measure how importantly the residual distribution changed with the Kolmogorov–Smirnov statistic for the residuals (J_r) themselves, which is defined in Figure 2. Of course, the two sets of residuals share measurement errors and logical blunders in the single set of arrival times on which they are based. Since the two samples are not independent, the confidence level (C_r) from the standard test of J_r underestimates the likelihood that the residual distribution changed. Nevertheless, computing C_r is a convenient way to normalize J_r for the number of residuals associated with each event.

In Middle America for almost every event where the residual distribution changed much at all, that is, for events where $C_r > 0.2$, the residuals improved for that individual event, $s_W{}^{3D} < s_W{}^{Bull}$ (Fig. 3, left). This is typical of most of the 12 regions where Chen and Willemann's relocation with

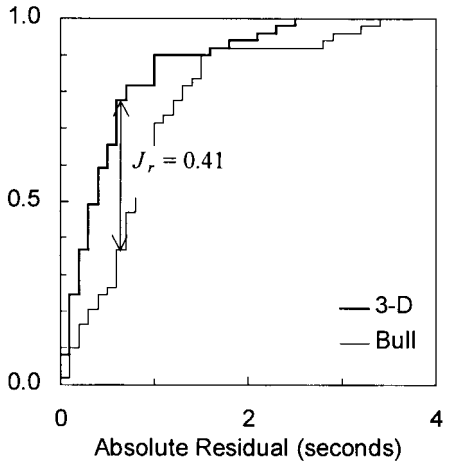

Figure 2. For an individual event, the difference between the distributions of residuals with respect to two alternative hypocenters can be measured by sorting the absolute values of the residuals then taking the maximum difference between their cumulative distribution functions, which is the Kolmgorov–Smirnov statistic J_r.

Kárason and van der Hilst's (1999) model significantly improved the dispersion of the residuals. In the southwestern Pacific, dispersion of the residuals became worse for more than one fifth of all events, in some cases even while changing the residuals importantly (Fig. 3, center). For each event in these regions where $s_W{}^{3D} > s_W{}^{Bull}$ and $C_r > 0.2$, however, the depth was fixed in the relocation. That is, the relocation failed to converge with a free depth, so the hypocenter was automatically fixed to the *Bulletin* depth for the event. This indicates that while the stations that contributed arrival times were not sufficiently well distributed to constrain hypocentral depth, the reported arrival times are inconsistent with the *Bulletin* depth if the Kárason and van der Hilst travel times are used.

In just one region, Pamir-Hindu Kush, $s_W{}^{3D} > s_W{}^{Bull}$ for a majority of the events (Fig. 3, right). For 18 of the 23 events where the new residuals are more widely dispersed, C_r is greater than 0.2. These earthquakes have free depths of 100–150 km in the *Bulletin*, and with the 3D model the hypocentral solutions converged with a free depth again, usually ending up 10–20 km deeper and occasionally more than 40 km deeper. Compared with most others in the *Bulletin*, an unusually large fraction of the arrivals associated with these events are from stations closer than 20°, many of them in Nepal. Station corrections were not used to compute either hypocenter or either set of residuals. Degradation of these fits suggests that when using data with a large proportion of regional arrival times, benefits will be realized from using better global travel time models only if variations in crustal thickness are taken into account.

Three-Way Comparison of Models

The principal evidence presented by Chen and Willemann (2001) for the importance of lateral variations in travel

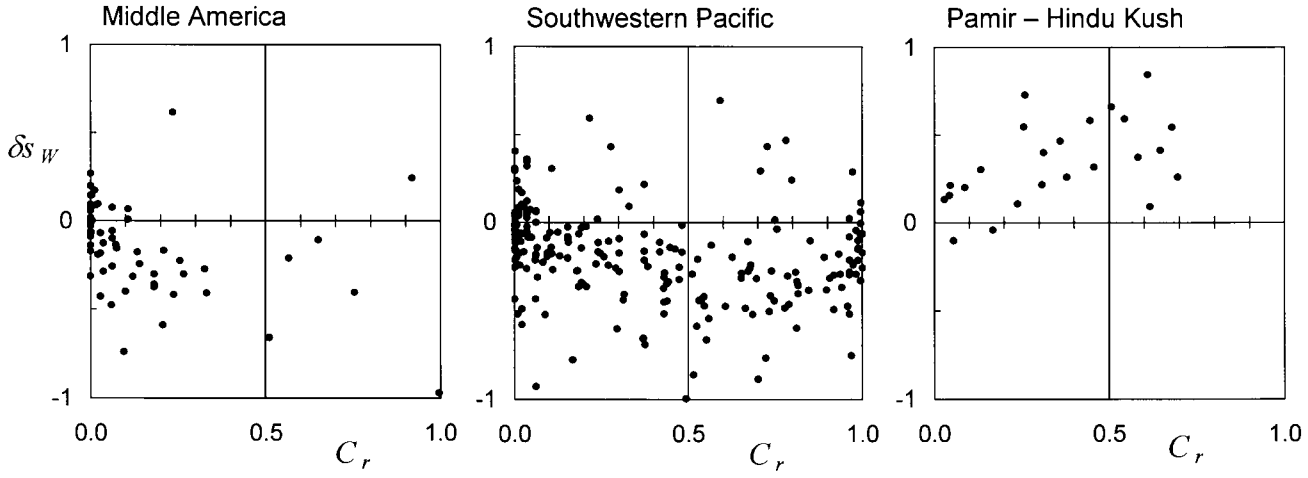

Figure 3. The effect and importance of relocation on residuals of an individual event can be summarized, respectively, by change in the rms of Winsorized residuals ($\delta s_W = s_W^{3D} - s_W^{Bull}$) and confidence that J_r is too large for the difference to arise by chance. *Left*: In the Middle America Flinn–Engdahl region, $s_W^{3D} < s_W^{Bull}$ for nearly all events where $C_r > 0.2$, which is typical of the Flinn–Engdahl regions where $C_s > 0.9$. *Center*: In Flinn–Engdahl regions 12–14 of the southwestern Pacific, events where $s_W^{3D} < s_W^{Bull}$ and $C_r > 0.2$ are more common, but for each of these anomalous events, depth was fixed during relocation due to an inadequate station distribution. *Right*: For most events in the Pamir–Hindu Kush Flinn–Engdahl region variance of the residuals grew worse as a result of relocating with a purportedly better earth model, even for events where J_r is large enough to suggest that the change was significant.

times is that Kárason and van der Hilst's tomographic model gave better results than ak135 for a relatively small number of reference events. For a full month of events, they recomputed hypocenters only using the tomographic model. The improvements in the arrival time residuals from these relocations are due both to a one-dimensional base model, ak135, that agrees better with observed travel times than the JB tables (Kennett *et al.*, 1995) and to lateral permutations in seismic wavespeeds from the tomographic inversion.

To determine what proportion of the improvement is due to each cause, I recomputed hypocenters for the same events with ak135 using the same relocation procedures described by Chen and Willemann (2001). These are essentially the procedures used in relocating events for the *Bulletin* and include iteratively computed uniform reduction weights. Among all relocated events around the world, the distribution of s_W improves about equally well going from the *Bulletin* to ak135, $J_s = 0.034$, $C_s = 97.9\%$, and from ak135 to the 3D model, $J_s = 0.037$, $C_s = 99.1\%$ (Fig. 4, left). These improvements are significant with high confidence even though they are smaller than improvements in some regions because of the larger number of events in the global sample.

Regionally, residuals for the ak135 hypocenters differ significantly from residuals for the JB hypocenters only in Flinn-Engdahl seismic regions New Zealand (Fig. 4, center), the Arctic, Pamir–Hindu Kush and Northern Eurasia (11, 40, 48, and 49). This is fewer than the approximately 5 of 48 regions where differences would be expected at the 90% confidence level by chance and, in addition, for earthquakes in Pamir-Hindu Kush residuals are significantly worse with either ak135 or the 3D model than with JB travel times. Thus, these results do not support a conclusion that there might be particular regions where ak135 is so much better than JB in reducing residuals for events typical of those in the *Bulletin* that ak135 might be adopted as a regional model.

On the other hand, residuals from the 3D model hypocenters differ significantly from residuals for the ak135 hypocenters only in seismic regions Alaska-Aleutians, Kermadec (Fig. 4, right), Fiji-Tonga and Vanuatu (1, 12, 13, and 14). These are all regions with strong lateral variations in wavespeed, but they are fewer than the number of regions expected to show improvement at the 90% confidence level by chance. Furthermore, improvement was not significant in other laterally heterogeneous regions such as Japan, South America, and the Sunda Arc. That is, the results do not support a conclusion that in regions of strong lateral heterogeneity the tomographic model produces an overwhelming reduction in arrival time misfits over any 1D model.

The other four Flinn–Engdahl seismic regions where changing from JB to the 3D model travel times improved residuals at the 90% confidence level were Middle America, Caroline Islands, Philippines, and Galapagos (5, 17, 22, and 44). But in these four regions neither changing from JB to ak135 travel times, nor changing from ak135 to the 3D model travel times, was alone sufficient to improve residuals at the 90% confidence level.

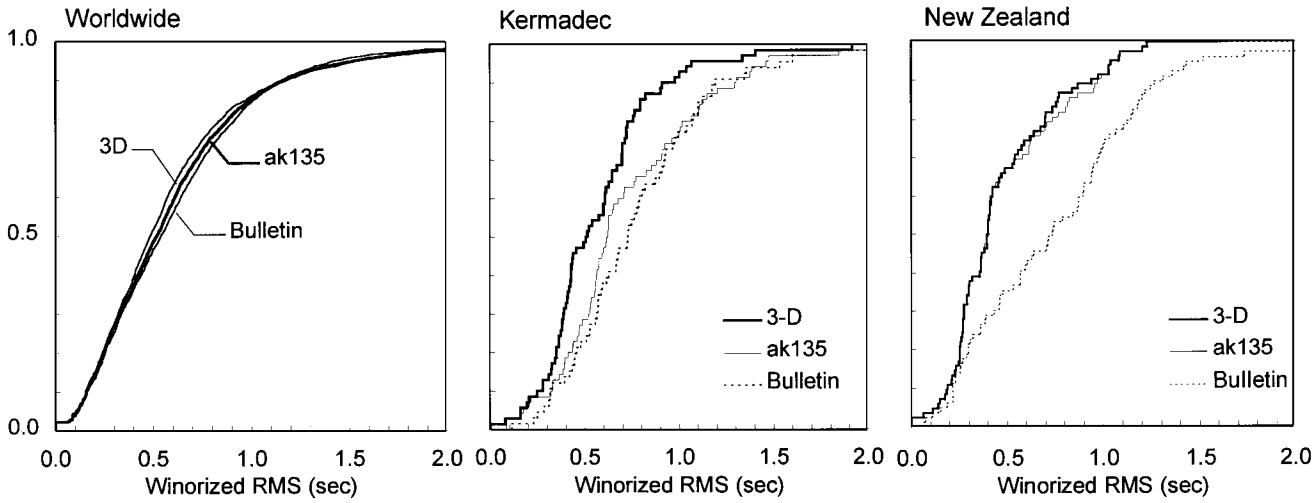

Figure 4. Residuals based on travel times from the J–B tables as in the *Bulletin*, from the 1D model ak135, and from a 3D tomographic model are compared using the cumulative distribution functions of the rms of the Winsorized residuals. *Left*: Among all events the differences are very small but, due to the size of the sample, statistically significant. *Center*: Along the Kermadec trench (Flinn–Engdahl region 12) where lateral heterogeneity is strong and there are no local stations, using a better 1D model provides only modest, statistically insignificant improvement, but using the 3D model offers much more improvement. *Right*: In New Zealand (Flinn–Engdahl region 11) where lateral heterogeneity is also strong, but there are many local stations, using a better 1D model very significantly improves residuals, and the 3D model provides no further improvement.

Conclusions

The fit to the data is an important feature of any set of hypocenters. Quantitative comparison of how well two sets of hypocenters fit the data is meaningful if the misfit measure is both robust and a measure of dispersion. Nevertheless, residuals from two sets of hypocenters computed from the same data are not really independent so measures of statistical significance must be interpreted carefully.

Despite limitations inherent in measuring arrival time misfit, an appropriate measure of how well observed times are fit can be a useful adjunct to other criteria for evaluating the accuracy of a new set of hypocenters. But it is imperative to adopt a misfit measure that is appropriate for comparing differently computed hypocenters, which is not necessarily the misfit measure that was best for computing either a travel time model or the hypocenters.

In the specific set of relocations used as examples of these statistical methods, only the combination of an improved 1D model and lateral variations clearly improves residuals compared with the JB travel times in a significant number of regions. Tests with more events might provide greater sensitivity and could conceivably demonstrate more significant improvement of residuals. Nevertheless, the limited improvement obtained in these examples suggests that errors in traditional arrival time picks are ubiquitous and large compared both with the residuals and with the corrections to residuals from 3D earth models.

Acknowledgments

This and all work at the ISC is partially supported by grants from 52 institutional members, including NSF Grant EAR97-25096. Peter Dawson and James Harris helped modify the ISC hypocenter location program and assisted with data management. Gary Pavlis provided a helpful review of the manuscript.

References

Anderson, K. R. (1982). Robust earthquake location using M estimates, *Phys. Earth Planet. Interiors* **30,** 119–130.

Buland, R. (1986). Uniform reduction error analysis, *Bull. Seism. Soc. Am.* **76,** 217–230.

Chen, Q.-f., and R. J. Willemann (2001), Relocations with 3-D *models, Bull. Seism. Soc. Am.* **91,** 1704–1716.

Engdahl, E. R., R. D. van der Hilst, and R. Buland (1998). Global teleseismic earthquake relocation with improved travel times and procedures, *Bull. Seism. Soc. Am.* **88,** 722–743.

Hollander, M., and D. Wolfe (1999). *Nonparametric Statistical Methods*, Wiley & Sons, New York, 770 pp.

Huber, P. J. (1996). *Robust Statistical Procedures*, Society for Industrial and Applied Mathematics, Philadelphia, 67 pp.

International Seismological Centre (ISC) (2000). *Bulletin of the International Seismological Centre*, Thatcham, Berkshire, U.K.

Jeffreys, H. (1932). An alternative to the rejection of observations, *Proc. R. Soc. Lond.* **187,** 78–87.

Jeffreys, H., and K. E. Bullen (1940). *Seismological Tables*, British Association for the Advancement of Science, London.

Karason, H., and R. D. van der Hilst (1999). New constraints on 3D variations in mantle P-wave speed, *EOS* **80,** F731.

Kennett, B. L. N., E. R. Engdahl, and R. Buland (1995). Constraints on

seismic velocities in the Earth from travel times, *Geophys. J. Int.* **122,** 108–124.

Pavlis, G. L. (1992). Appraising relative earthquake location errors, *Bull. Seism. Soc. Am.* **82,** 836–859.

Ritzwoller, M. H., M. P. Barmin, A. Villasenor, A. L. Levshin, E. R. Engdahl, W. Spakman, and J. Trampert (2000). Construction of a 3-D P and S model of the crusts and upper mantle to improve region alocation in W. China, Central Asia and parts of the Middle East, in *21st Seismic Research Symposium*, Los Alamos National Laboratory, Los Alamos, New Mexico, LA-UR-99-4700. 656–665.

Sultanov, D. D., J. R. Murphy, and K. D. Rubinstein (1999). A seismic source summary for Soviet peaceful nuclear explosions, *Bull. Seism. Soc. Am.* **89,** 640–647.

Thurber, C., F. Haslinger, and C. Trabant (2000). Testing event location capability with ground truth events in Kazakstan, in *21st Seismic Research Symposium*, Los Alamos National Laboratory, Los Alamos, New Mexico, LA-UR-99-4700. 283–293.

Thurber, C. H., and E. R. Engdahl (2000). Advances in global seismic event location, in *Advances in Seismic Event Location*, C.H. Thurber and N. Rabinowitz (Editors), Kluwer Academic Publishers, Dordrecht, the Netherlands, 3–22.

Wilcox, R. R. (1997). *Introduction to Robust Estimation and Hypothesis Testing*, Academic Press, San Diego, 352 pp.

International Seismological Centre
Pipers Lane
Thatcham, Berkshire RG19 4NS
United Kingdom
ray@isc.ac.uk.

Manuscript received 11 January 2002.