Supplementary

Integrated Earthquake Catalog of the Ossetian Sector of the Greater Caucasus

Method

Modification of Nearest Neighbor Method

At the input there are two catalogs, the main Catalog 1 and additional Catalog 2. We believe that neither Catalog 1 nor Catalog 2 contains duplicates within themselves. The problem is to find records in Catalog 1 for records in Catalog 2 that will correspond to the same seismic events (duplicates) and divide Catalog 2 into events that have duplicates in Catalog 1 and unique events. The two-step modification of the nearest neighbor method (Figure S1) is based on the following provision: duplicates form pairs, in which events necessarily belong to different source catalogs.

Step I. For each event of the additional Catalog 2, we look for the nearest neighbor from the main Catalog 1 in accordance with the chosen metric. This step is similar to the classic nearest neighbor method. Thus, for each event from Catalog 2, a single event from Catalog 1 is determined for which it can be a duplicate.

Step II. Some events from the main catalog 1 may occur to be closest for several events from additional catalog 2. This is shown in Figure S1, where several gray dots (earthquakes from additional catalog) are associated with the same blue dot (earthquake from main catalog). This case the closest of such events is selected as a potential duplicate. This is illustrated in Figure S1 by red arrow. Other events are declared to be non-duplicates, regardless of the metric values.

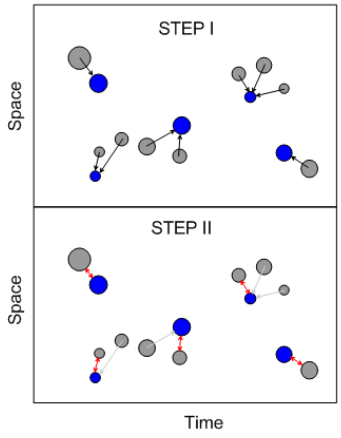

Figure S1. Modification of the nearest neighbor method for identifying duplicates in earthquake catalogs. Blue and gray circles are events of the main and additional catalogs, respectively, and identified pairs are shown by red arrows on the Step II panel.

After the second step, the nearest neighbors are not defined for all events, because some events from the Catalog 1 were not closest to any event from the Catalog 2 and vice versa. However, there may be duplicates among them. We exclude from the analysis the events of the first and second catalogs, which found their pair at the first stage. For the remaining events, we again define pairs. The procedure is repeated until all events from the catalog with a smaller number of events find their pair. At the same time, some of the pairs are not actually duplicates. Therefore, a threshold value is introduced for the neighborhood function.

As a result, we consider the events of additional Catalog 2 with the value of the neighborhood function less than the threshold one as duplicates. The rest of Catalog 2 events are declared unique and added to Catalog 1. Selection of neighborhood function and threshold determination are discussed below. Further, any number of catalogs can be sequentially added. One of the advantages of the described method is the predetermined priority of data sources. The procedure ensures that events with a higher priority are automatically included in the final catalog.

Neighborhood Function for Duplicates in Earthquake Catalogs: A Probabilistic Approach

We rely on a probabilistic model, in our task in the station error space. It is assumed that the difference in earthquakes identified by different networks has a normal distribution for each of the parameters:

$$f(DT) = \frac{1}{\sigma_T\sqrt{2\pi}}\exp(-\frac{(DT-\overline{DT})^2}{2{\sigma_T}^2});$$
$$f(DX) = \frac{1}{\sigma_X\sqrt{2\pi}}\exp(-\frac{(DT-\overline{DT})^2}{2{\sigma_T}^2}) \quad (1);$$
$$f(DY) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp(-\frac{(DY-\overline{DY})^2}{2{\sigma_Y}^2})$$

Here *DT, DX, DY*, are the difference in time, longitude and latitude between the nearest events from the main and additional catalogs; σ_T, σ_X, σ_Y, are the corresponding standard deviations, and and $\overline{DT}$, $\overline{DX}$, $\overline{DY}$, are average values. We assume that all differences are independent random variables, then the probability of duplicate will be the product of the probabilities of individual parameters. This will be the multivariate normal distribution:

$$f(DX,DY,DT) = \frac{1}{\sigma_T\sigma_X\sigma_Y(2\pi)^{3/2}}\cdot\exp\left(-\left(\frac{(DT-\overline{DT})^2}{2{\sigma_T}^2}+\frac{(DX-\overline{DX})^2}{2{\sigma_X}^2}+\frac{(DY-\overline{DY})^2}{2{\sigma_Y}^2}\right)\right) \quad (2)$$

Thus, we naturally arrive at the Euclidean metric.

$$Ro = \sqrt{\frac{(DT-\overline{DT})^2}{{\sigma_T}^2}+\frac{(DX-\overline{DX})^2}{{\sigma_X}^2}+\frac{(DY-\overline{DY})^2}{{\sigma_Y}^2}} \quad (3)$$

If the standard deviations of the parameters is correctly determined, the metric can be easily converted into the probability of a given pair to be a duplicate. It is easy to see that metric (3) is simply the radius of the sphere, measured in standard deviations. The metric allows to take into account the systematic bias and dispersion of each of the parameters. In practice, it makes sense to take into account the systematic bias if it has a value of the order of standard deviation or more.

Dtermination of the threshold value of the metric Ro separating dudlicates and unique events

At each stage, the merging of catalogs is performed in two steps. First, parameters of the metric (3) are determined. For this, the metric (3) Ro between the nearest events from two source catalogs with standard parameters $\sigma_T = 0.05$ min, $\sigma_X = \sigma_Y = 15$ km is calculated. The threshold value of the metric $Ro = 10$ is used for the preliminary duplicate identification, which corresponds to the time and space difference of 0.5 min / 150 km. From our experience, we know that the above numerical parameters are characteristic of most instrumental catalogs. Standard deviations σ_T, σ_X, and, σ_Y of the variables *DT*, *DX*, and *DY* are calculated for the pre-identified duplicates. The illustration of the first step is presented in Figures 4,6,8 of the main text.

In the next step, a threshold value of the metric is determined and the final duplicate identification takes place.To determine the metric threshold, we calculate the probability of missing duplicates (error of the first kind) and the probability of false duplicates (error of the second kind). We assume that the maximum value of the metric for events that can be duplicates is $Ro = 30$, which corresponds to the time and space difference of about 1.5 min and about 500 km. We construct the distribution of such events Fdub. The red lines in Figures 5b, 7b, 9b represent the

value of $1 - F$dub, which we consider as the probability of missing a duplicate (error of the first kind). The probability of a false duplicate (error of the second kind) is estimated as follows: we calculate the values of the metric (3) Ro between events within an additional catalog. The blue lines in Figures 5b, 7b, 9b represent the proportion of events with a proximity less than a given value of Ro. The black lines show the estimate of the total probability of the first and the second kind errors. The black lines show the estimate of the total probability of the first and the second kind errors. The threshold value of the metric minimizes the total number of errors, it is shown by dashed line. Figures 5c, 7c, 9c show the distribution of normalized times and distances for the nearest events from the merged catalogs. The metric contour lines correspond to the chosen threshold value, which provides optimal separation of duplicates and unique events.

Before the merging process, each of the source catalogs (Table 1) was checked for internal duplicates. The statistical analysis did not reveal any anomalous groups of close events (Figure S1). If there is a noticeable number of internal duplicates in the catalog, the distribution of the proximity function (*Ro* metric) between the nearest events has a characteristic bimodal shape with a minimum between modes at $Ro \approx 10$ [1,4–6]. For all source catalogues, there is no mode of metric distribution in small values. The number of close events with a metric value $Ro < 10$ is very small, less than 0.5%. We do not consider these events to be duplicates because, due to natural clustering, earthquakes can occur very close in space and time. The analysis (Figure S1) was performed with metric parameters $\sigma_T = 0.05\ min$, $\sigma_X = \sigma_Y = 15\ km$.

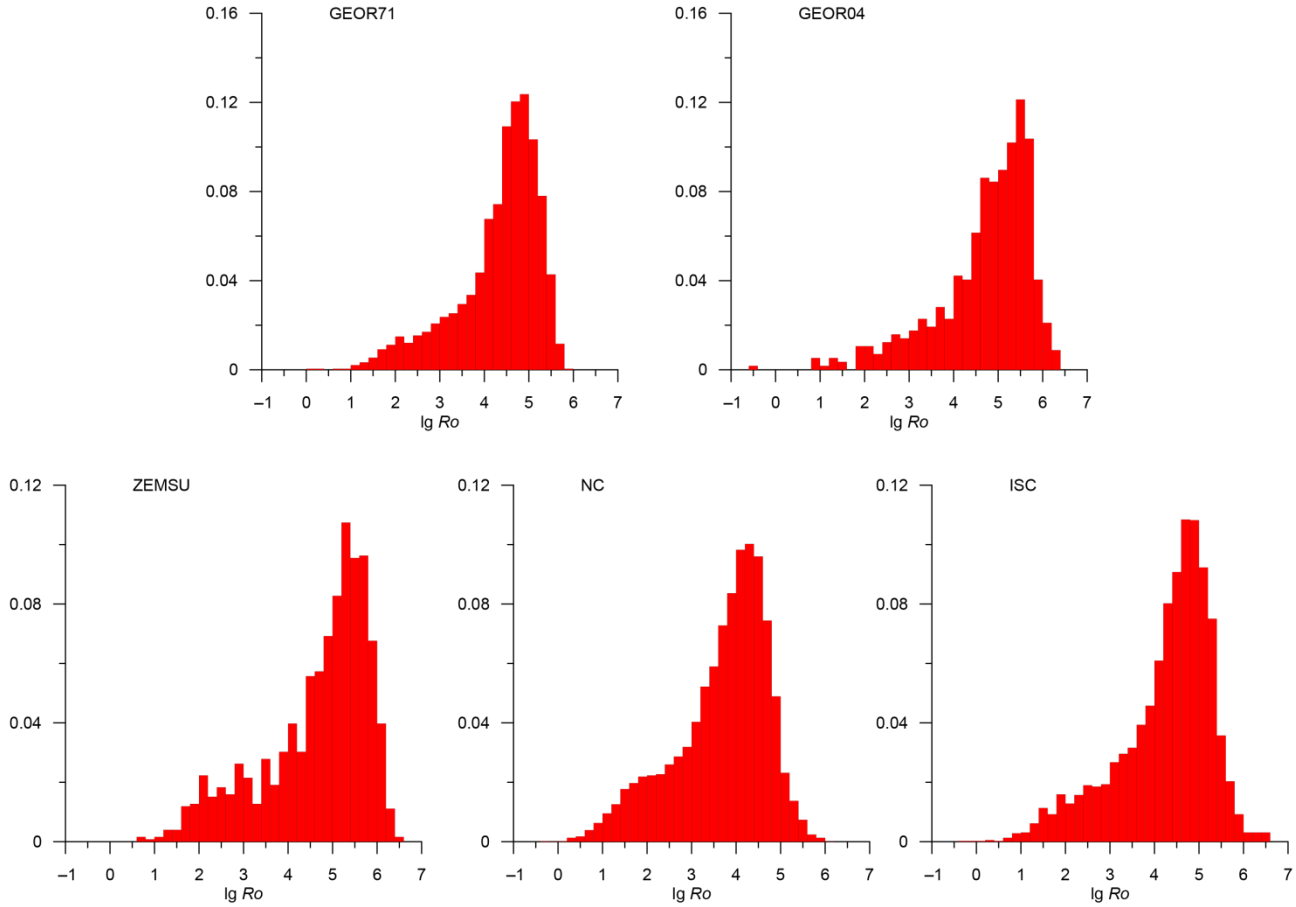

Figure S2. Distribution of the metric for events within the source earthquake catalogs (Table 1). The catalog name is indicated on the histogram.

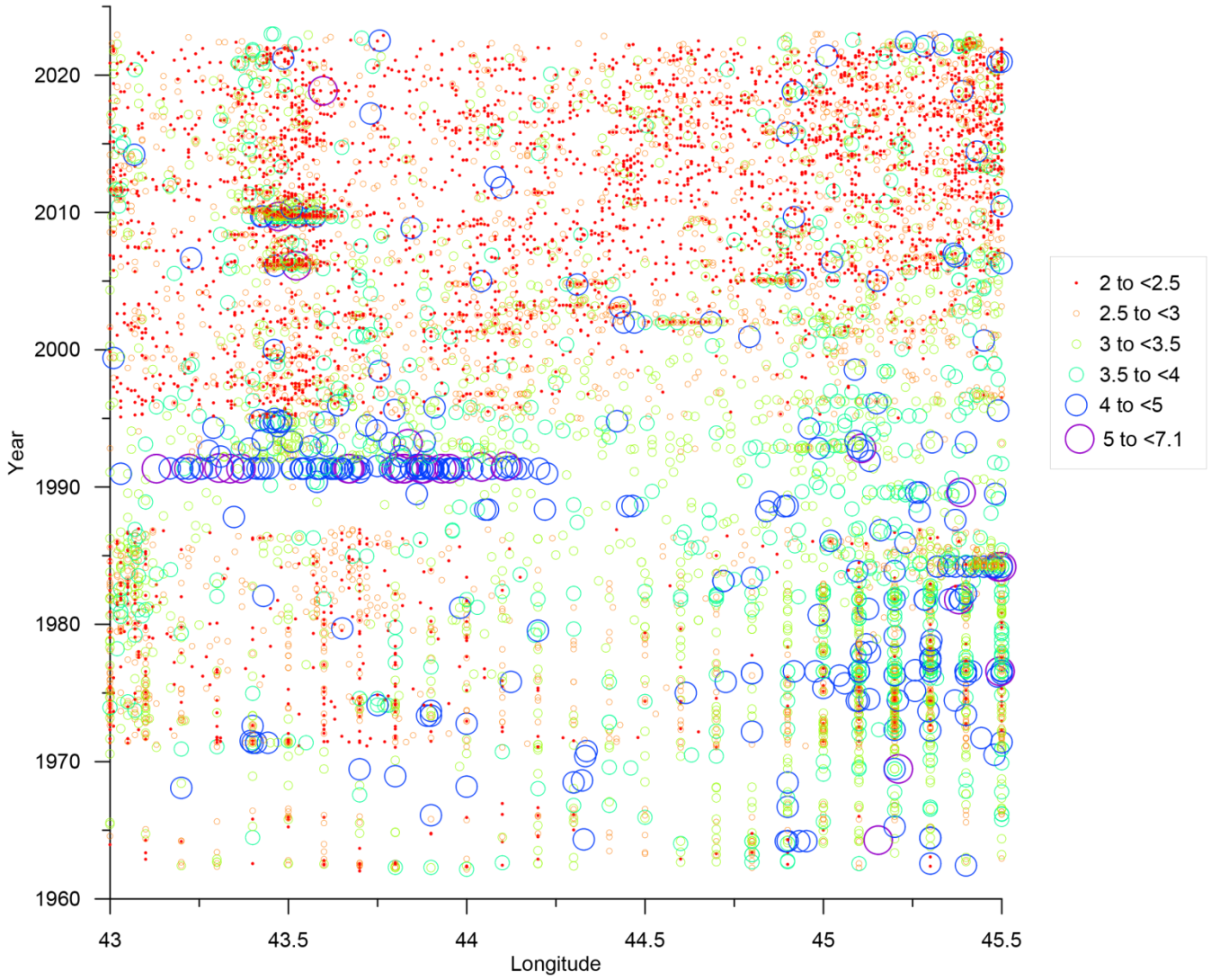

Figure S3. Distribution of earthquake magnitudes in space and time.