
https://doi.org/10.1093/gji/ggac333

Institutional Repository GFZpublic: https://gfzpublic.gfz-potsdam.de/
Effects of large aftershocks on spatial aftershock forecasts during the 2017–2019 western Iran sequence

Behnam M. Asayesh,1,2,3 Hamid Zafarani,1 Sebastian Hainzl3 and Shubham Sharma2,3

1 Seismological Research Center, International Institute of Earthquake Engineering and Seismology (IIEES), 19537-14453 Tehran, Iran. E-mail: maleki@gfz-potsdam.de
2 Institute of Environmental Sciences and Geography, University of Potsdam, 14476 Potsdam, Germany
3 Geophysics, GFZ German Research Centre for Geosciences, 14467 Potsdam, Germany

Accepted 2022 August 17. Received 2022 August 17; in original form 2022 January 1

SUMMARY
On 12 November 2017, an earthquake with a moment magnitude of 7.3 struck the west of Iran near the Iraq border. This event was followed about 9 and 12 months later by two large aftershocks of magnitude 5.9 and 6.3, which together triggered intensive seismic activity known as the 2017–2019 Kermanshah sequence. In this study, we analyse this sequence regarding the potential to forecast the spatial aftershock distribution based on information about the main shock and its largest aftershocks. Recent studies showed that classical Coulomb failure stress (CFS) maps are outperformed by alternative scalar stress quantities, as well as a distance-slip probabilistic model (R) and deep neural networks (DNN). In particular, the R-model performed best. However, these test results were based on the receiver operating characteristic (ROC) metric, which is not well suited for imbalanced data sets such as aftershock distributions. Furthermore, the previous analyses also ignored the potential impact of large secondary earthquakes. For the complex Kermanshah sequence, we applied the same forecast models but used the more appropriate MCC-F1 metric for testing. Similar to previous studies, we also observe that the receiver independent stress scalars yield better forecasts than the classical CFS values relying on the specification of receiver mechanisms. However, detailed analysis based on the MCC-F1 metric revealed that the performance depends on the grid size, magnitude cut-off and test period. Increasing the magnitude cut-off and decreasing the grid size and period reduce the performance of all methods. Finally, we found that the performance of the best methods improves when the source information of large aftershocks is additionally considered, with stress-based models outperforming the R model. Our results highlight the importance of accounting for secondary stress changes in improving earthquake forecasts.

Key words: Earthquake hazards; Earthquake interaction; forecasting; and prediction; Statistical seismology.

1 INTRODUCTION
Large earthquakes almost always trigger subsequent earthquakes, so-called aftershocks, which make up an aftershock sequence that can take months to years. Sometimes those aftershocks can be as destructive or deadly as the main shock or even worse, for example the 2011 Mw 6.2 Christchurch event, triggered by the Mw 7.1 Darfield main shock (Stramondo et al. 2011). Therefore, immediately after a large earthquake, an accurate probabilistic forecast of the spatial and temporal aftershock distribution is of utmost importance for planning rescue activities, emergency decision-making and risk mitigation in the disaster area.

Large earthquakes cause permanent deformations of the surrounding ground and permanently alter the stresses on nearby faults. These stress changes can trigger aftershocks, whose temporal decay is well described by the Omori law (Utsu 1961), but explaining and forecasting their spatial distribution is more difficult. Based on many examples showing a clear correlation between static Coulomb failure stress (CFS) changes and the spatial aftershock distribution, the static stress transfer hypothesis is perhaps the most widely accepted and used criterion to explain aftershocks triggering and their spatial pattern (Harris & Simpson 1992; King et al. 1994; Toda et al. 1998; Parsons et al. 1999; Asayesh et al. 2020; Jamalreyhani et al. 2020).

According to the CFS hypothesis, aftershocks concentrate at the sites with increased stress (ΔCFS > 0), while earthquake activity is suppressed in areas where stress changes are negative (ΔCFS < 0), so-called stress shadows (Harris & Simpson 1992). However,
the applicability of this hypothesis is still ambiguous because of the missing effect of dynamic stress triggering (Felzer & Brodsky 2006) and the lack of clear observational evidence for stress shadows (Harris & Simpson 2002; Felzer & Brodsky 2005). Furthermore, ΔCFS calculations suffer from large uncertainties such as the non-uniqueness of slip inversions (Hainzl et al. 2009), secondary stress triggering (Helmstetter et al. 2005) and unknown receiver fault mechanisms. The two principal approaches for ΔCFS calculations are to assume a specified receiver fault or an optimally oriented plane (OOP), respectively (King et al. 1994; Lin & Stein 2004; Ma et al. 2005; Toda 2008; Ishitbe et al. 2011). Both approaches suffer from unknowns such as poorly constrained fault geometries, unrecognized blind faults, and an uncertain background stress field. To account for these problems, Steacy et al. (2005) suggested using the regional fault trends to fix the fault strike and only optimize dip and rake, while Hainzl et al. (2010) proposed to calculate ΔCFS on distributions of possible receiver fault orientations. Recent analysis of the M 7.2, 2010 El Mayor-Cucapah sequence by Segou & Parsons (2020) showed that aftershocks seldom occur on OOPs (35 per cent of the total events). On the other hand, resolved total stress (coseismic + background) on all possible rupture planes with constraints taken from past earthquake ruptures could provide a physical explanation for 89 per cent of analysed aftershocks during this sequence.

Recently, some studies using ROC analysis have further questioned the ability of ΔCFS values in forecasting aftershock locations. Performing a thorough analysis of various scalar metrics, Meade et al. (2017) showed that a number of alternative scalar metrics, which do not need any specification of the receiver mechanism, are better predictors of aftershock locations than ΔCFS for fixed receiver orientation. Those superior stress scalars comprise the maximum shear, the von-Mises yield criterion (a scaled version of the second invariant of the deviatoric stress-change tensor), and the sum of the absolute values of the independent components of the stress-change tensor. Furthermore, DeVries et al. (2018) showed that deep neural network (DNN) techniques also clearly outperform the ΔCFS-based forecasts. Mignan & Broccardo (2019) demonstrated that even a single neuron provided results similar to the complex DNN and proposed a model for the aftershock distribution, which only depends on the mean slip of the main shock and the distance to the rupture plane. The demonstrated success of this simple model further questioned the usefulness of stress-tensor calculations. Later, Sharma et al. (2020) challenged all previous results by using all available slip inversions from SCRMOD database (Mai & Thingamjam 2014) and repeating the analysis with more appropriate stress calculations. In particular, they considered layered-half spaces and variable receiver mechanisms as well as alternative tests, finding that ΔCFS resolved on OOPs or calculated for variable mechanism (VM) significantly improved the ΔCFS results. However, the maximum shear (MS) and von-Mises stress (VMS) scalar stress metrics performed still better than all receiver-dependent metrics. They also showed that the binary forecasts of the distance-slip (R) model were best for forecasting the aftershock area. All studies mentioned above only considered the main shock source for calculating the stress scalars or the distance-slip value. They ignored secondary static stress changes caused by aftershocks. However, several other studies suggest that smaller but more frequent earthquakes can locally play an equally or even more important role in redistributing stress (Michael & Jones 1998; Felzer et al. 2002; Helmstetter 2003; Helmstetter et al. 2005; Meier et al. 2014; Nandan et al. 2016; Cattania et al. 2018; Mancini et al. 2019; Asayesh et al. 2020b). Therefore, in addition to the stress induced by the main shock, the stress changes related to large aftershocks should be considered in forecasting the rate and location of future aftershocks.

Another weak point of previous tests is the use of ROC curves because ROC analysis can be misleading for imbalanced data (Davis et al. 2005; Davis & Goadrich 2006; Jeni et al. 2013; Saito & Rehmsmeier 2015). Parsons (2020) argued that the ROC tests suffer from data imbalance because most areas lack aftershocks (many more negative cases than positive), inhibiting any resolving power of the ROC analysis for aftershock data. In order to address this drawback, Cao et al. (2020) recently proposed the MCC-F1 curve that combines two informative single-threshold metrics, the Matthews correlation coefficient (MCC) and the F1 score. The MCC-F1 curve largely solves the misleading conclusions for imbalanced data sets by providing clearer and more consistent way to evaluating binary classifiers. Furthermore, it provides a useful means to identify the best threshold for classification (Cao et al. 2020).

In this study, we analysed the Kermanshah (west Iran) 2017–2019 sequence using the same models but additionally accounting for the effect of significant aftershocks and using the MCC-F1 test metric. The Kermanshah sequence took place in Iranian plate which has frequently experienced catastrophic earthquakes resulting in massive loss of lives, mass homelessness, and disruptions of agricultural and industrial lifelines. This sequence started with the 2017 M7.3 Azgeleh main shock (also termed in other studies Erzgeleh, Kermanshah, Darbandikhan, Saropol-e Zahab, or Erzgeleh-Saropolzahab), followed by 17 events with magnitude equal to or larger than 5. The well-recorded aftershocks and available slip models for the main shock and the largest aftershocks enabled us to explore the role of secondary slip and stress changes using different binary tests (ROC and MCC-F1). Furthermore, we analysed the dependence of the results on the chosen magnitude cut-off, grid size and aftershock sequence duration.

2 THE 2017–2019 KERMANSHAH SEQUENCE

2.1 Seismotectonic setting

Active faulting and folding, volcanic activities, mountainous terrain and variable crustal thickness are characteristics of the Iranian plateau, resulting from the convergence of Arabian and Eurasian plates (20–30 mm yr⁻¹; Vernant et al. 2004). This convergence is accommodated across different seismotectonic zones such as Zagros, Alborz, Central Iranian blocks and other seismotectonic zones in the NW and NE of Iran. The Zagros range in southwestern Iran, as a central part of the Alpine-Himalayan orogenetic belt (inset map in Fig. 1), accommodates approximately one-third to one-half of the current Arabian-Eurasian convergence (Vernant et al. 2004; Masson et al. 2005). The seismicity in Zagros generally consists of low to moderate magnitude events (Barnhart et al. 2013; Karasözen et al. 2017) distributed across a ~200-km-wide zone. These earthquakes commonly occur on blind faults (Berberian 1995), located beneath the Hormuz Salt Formation within the sedimentary cover and basement between 8 and 14 km (Jackson & Fitch 1981; Nî & Barazangi 1986; Baker et al. 1993; Hessami et al. 2001; Jamalreyhani et al. 2021).

2.2 Earthquake activity

The 2017–2019 Kermanshah (west Iran) sequence started on 12 November 2017 (18:18 UTC) with a main shock known as the Azgeleh earthquake that occurred in Zagros fold-and-thrust belt
Figure 1. Main tectonic features of the Iranian plateau and the Kermanshah area, west of Iran. Epicentral and depth distribution of the main shocks (Azgeleh, Tazehabab and Sarpol-e Zahab earthquakes) (big stars) and their following seismicity (circles and small stars) from Iranian Seismological Center (IRSC). The stars show the events with a magnitude $M_w \geq 5$. The colour of stars and circles depict time. The black bitch balls show the focal mechanism of main shocks (Fathian et al. 2021). The big green squares show the big cities, and small green squares depict the main towns near the epicentral area of the main shocks (Az: Azgeleh, TA: Tazehabad, SZ: Sarpol-e Zahab and QS: Qasr-e Shirin). The causative faults of the main shocks are depicted by abbreviations (AzF Listric: surface projection of listric slip model, TAF: Tazehabad fault and SPF: Sarpol-e-Zahab fault). The black solid and dashed lines in the map and inset maps h o w the main active faults (MFF: Main Front Fault, HZF: High Zagros Fault and KhF: Khanaqin Fault), and black triangles are the IRSC stations. The dashed red line encircles the selected events in this study.

(ZFTB) near the Iraq border. Such a severe main shock with a moment magnitude of 7.3, damaging a region between the towns of Azgeleh to Sarpol-e Zahab, were never documented before over the last several centuries (Zafarani et al. 2020). With its about 630 casualties, immense building damage, and economic losses (Ahmadi & Bazargan-Hejazi 2018), it also was the most catastrophic earthquake in Iran over the last decade. Based on a comprehensive study by Fathian et al. (2021), the event had an oblique-slip thrust mechanism with a considerable dextral component (Strike: 354°, Dip: 16° and Rake: 137°). Approximately nine months later, this main shock was followed by a moderate event ($M_w 5.9$) on 25 August 2018 (22:13 UTC) that occurred 45 km to the east near the Tazehabad city and killed three inhabitants. Reported focal mechanisms from national and international seismological agencies and interferometric observation revealed that this event was a left-lateral strike-slip event (Strike: 267°, Dip: 78° and Rake: 2°) on an unknown fault (Fathian et al. 2021). The third large event, the so-called Sarpol-e Zahab earthquake, occurred about 1 yr after the Azgeleh main shock on 25 November 2018, 16:37 UTC, with a magnitude $M_w 6.3$. It had a right-lateral strike-slip mechanism (Strike: 34°, Dip: 63° and Rake: 170°) and occurred near the Sarpol-e Zahab and Qasr-e Shirin towns (Figs 1 and S1–S3).
Figure 2. (a) Magnitude versus time plot of events occurred within a distance of 75 km to the Azgeleh main shock rupture with depth less than 20 km during the 2017–2019 Kermanshah sequence. Stars show the largest earthquake ($M_w \geq 5$). (b) Noncumulative frequency–magnitude distribution, where the completeness magnitude ($M_c$) is estimated as 2.2 according to the maximum curvature (MaxC) method, while the fit of the Ogata and Katsura model (solid lines) yields 2.5 and 2.8 for detection probabilities of 90 and 99 per cent (dotted horizontal lines), respectively. (c) Time-series of the main shock and its largest aftershocks ($M_w \geq 5$). Here the numbers in brackets between events denote the number of $M \geq 1.0$ events recorded in the interevent times.

Table 1. Characteristics of the 2017–2019 Kermanshah sequence main shock and 17 following large events with magnitude $\geq 5$.

<table>
<thead>
<tr>
<th>No.</th>
<th>Date</th>
<th>Time</th>
<th>Lat.</th>
<th>Lon.</th>
<th>Depth</th>
<th>Strike</th>
<th>Dip</th>
<th>Rake</th>
<th>$M_w$</th>
<th>Slip model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.11.17</td>
<td>18:18</td>
<td>34.730</td>
<td>45.762</td>
<td>18.1</td>
<td>354$^b$</td>
<td>16$^b$</td>
<td>137$^b$</td>
<td>7.3</td>
<td>b</td>
</tr>
<tr>
<td>2</td>
<td>11.12.17</td>
<td>14:09</td>
<td>35.078</td>
<td>45.840</td>
<td>8.0</td>
<td>133$^c$</td>
<td>71$^c$</td>
<td>26$^c$</td>
<td>5.5</td>
<td>d and e</td>
</tr>
<tr>
<td>3</td>
<td>06.01.18</td>
<td>07:14</td>
<td>33.722</td>
<td>45.709</td>
<td>8.4</td>
<td>346$^c$</td>
<td>50$^c$</td>
<td>106$^c$</td>
<td>5.5</td>
<td>d, e, f, and g</td>
</tr>
<tr>
<td>4</td>
<td>01.02.18</td>
<td>16:37</td>
<td>34.359</td>
<td>45.712</td>
<td>10.5</td>
<td>34$^b$</td>
<td>63$^b$</td>
<td>170$^b$</td>
<td>6.3</td>
<td>b</td>
</tr>
<tr>
<td>5</td>
<td>11.01.18</td>
<td>17:09</td>
<td>34.348</td>
<td>45.665</td>
<td>7.2</td>
<td>336$^c$</td>
<td>26$^c$</td>
<td>111$^c$</td>
<td>5.3</td>
<td>b</td>
</tr>
<tr>
<td>6</td>
<td>11.01.18</td>
<td>22:17</td>
<td>33.709</td>
<td>45.699</td>
<td>8.8</td>
<td>341$^c$</td>
<td>37$^c$</td>
<td>109$^c$</td>
<td>5.0</td>
<td>d and e</td>
</tr>
<tr>
<td>7</td>
<td>01.02.18</td>
<td>08:00</td>
<td>33.724</td>
<td>45.686</td>
<td>10.9</td>
<td>348$^c$</td>
<td>43$^c$</td>
<td>108$^c$</td>
<td>5.3</td>
<td>d, e, f, and g</td>
</tr>
<tr>
<td>8</td>
<td>11.01.18</td>
<td>12:17</td>
<td>33.709</td>
<td>45.699</td>
<td>8.8</td>
<td>341$^c$</td>
<td>37$^c$</td>
<td>109$^c$</td>
<td>5.0</td>
<td>d and e</td>
</tr>
<tr>
<td>9</td>
<td>01.02.18</td>
<td>08:35</td>
<td>33.709</td>
<td>45.699</td>
<td>8.8</td>
<td>341$^c$</td>
<td>37$^c$</td>
<td>109$^c$</td>
<td>5.0</td>
<td>d and e</td>
</tr>
<tr>
<td>10</td>
<td>11.01.18</td>
<td>14:09</td>
<td>35.078</td>
<td>45.840</td>
<td>8.0</td>
<td>133$^c$</td>
<td>71$^c$</td>
<td>26$^c$</td>
<td>5.5</td>
<td>d and e</td>
</tr>
<tr>
<td>11</td>
<td>22.07.18</td>
<td>10:07</td>
<td>34.630</td>
<td>46.255</td>
<td>8.0</td>
<td>246$^e$</td>
<td>83$^e$</td>
<td>-19$^e$</td>
<td>5.8</td>
<td>d and e</td>
</tr>
<tr>
<td>12</td>
<td>25.08.18</td>
<td>22:13</td>
<td>34.622</td>
<td>46.233</td>
<td>8.3</td>
<td>267$^b$</td>
<td>78$^b$</td>
<td>2$^b$</td>
<td>5.9</td>
<td>b</td>
</tr>
<tr>
<td>13</td>
<td>25.11.18</td>
<td>16:37</td>
<td>34.359</td>
<td>45.712</td>
<td>10.5</td>
<td>34$^b$</td>
<td>63$^b$</td>
<td>170$^b$</td>
<td>6.3</td>
<td>b</td>
</tr>
<tr>
<td>14</td>
<td>25.11.18</td>
<td>17:09</td>
<td>34.348</td>
<td>45.665</td>
<td>7.2</td>
<td>336$^c$</td>
<td>26$^c$</td>
<td>111$^c$</td>
<td>5.3</td>
<td>b</td>
</tr>
<tr>
<td>15</td>
<td>26.11.18</td>
<td>00:38</td>
<td>34.379</td>
<td>45.742</td>
<td>7.6</td>
<td>7$^c$</td>
<td>80$^c$</td>
<td>170$^c$</td>
<td>5.0</td>
<td>b</td>
</tr>
<tr>
<td>16</td>
<td>06.01.19</td>
<td>13:41</td>
<td>34.151</td>
<td>45.648</td>
<td>10.9</td>
<td>351$^c$</td>
<td>30$^c$</td>
<td>124$^c$</td>
<td>5.7</td>
<td>d and e</td>
</tr>
<tr>
<td>17</td>
<td>01.04.19</td>
<td>10:07</td>
<td>33.717</td>
<td>45.697</td>
<td>16.3</td>
<td>323$^c$</td>
<td>36$^c$</td>
<td>99$^c$</td>
<td>5.3</td>
<td>d and e</td>
</tr>
<tr>
<td>18</td>
<td>11.05.19</td>
<td>10:28</td>
<td>34.870</td>
<td>45.755</td>
<td>8.3</td>
<td>0$^c$</td>
<td>50$^c$</td>
<td>98$^c$</td>
<td>5.1</td>
<td>d and e</td>
</tr>
</tbody>
</table>

$^a$Iranian Seismological Center (IRSC), $^b$Fathian et al. (2021), $^c$Global Centroid Moment Tensor Catalog (GCMT), $^d$Wells & Coppersmith (1994), $^e$Hanks & Kanamori (1979), $^f$Barnhart et al. (2018) and $^g$Nissen et al. (2019).

These three earthquakes triggered intense seismicity within their epicentral area. About 8000 earthquakes were recorded by the Iranian Seismological Center (IRSC) during 600 d after the Azgeleh main shock (Fig. 2a). Out of these aftershocks, about 3000 events had a magnitude equal to or greater than 2.5. In addition to the largest three events, the recorded sequence consists of 15 $M_w \geq 5$ events (Table 1 and Fig. 2c) and about 150 events with magnitudes in the range of 4 $\leq M < 5$.

For our further analysis, we selected all events that occurred during the first 600 d (about 20 months) until 4 July 2019, with depth.
less than 20 km and a horizontal distance of less than 75 km to the Azgeleh main shock rupture (dashed red line in Fig. 1). For choosing this target region, we just considered the surface and depth distribution of the observed earthquake. The events’ depths were mainly less than 20 km and concentrated inside an area less than 75 km to the main shock rupture. Thus, we used these limits to define the target volume. This data set consists of 7800 aftershocks recorded by IRSC with a magnitude range of 1.0–6.3. Although some of these events were probably not related to the main shock, we consider all of them as aftershocks. More than 1000 of these aftershocks were detected and located by a dense local seismological network around the epicentral area of the main shock (Fathian et al. 2021) and ~700 events with magnitude equal or greater than 3 have been relocated by (Jamalreyhani et al. 2020b). All together the location error is 3 km. Fig. 2 shows the temporal occurrence and the frequency–magnitude distribution of the selected events. We estimated the completeness magnitude ($M_c$) and the frequency–magnitude distribution of the selected events.

We estimated the completeness magnitude ($M_c$) of the data set by two different standard procedures. First, the fast and straightforward ‘maximum curvature method’ (Wyss et al. 1999; Wiemer & Wyss 2000) defines $M_c$ by the maximum of the magnitudes’ histogram, that is the maximum value of the first derivative of the cumulative frequency–magnitude distribution. This method yields a completeness magnitude of 2.2. Secondly, we used the Ogata & Katsura (1993) model that considers a detection probability function to fit the entire magnitude range (EMR). The application of the latter approach yields an $M_c$ value of 2.5 and 2.8 for a 90 and 99 per cent confidence level, respectively (Fig. 2b).

2.3 Earthquake source information

For the aftershock forecasts based on stress metrics or R-model values, appropriate source models for the main shock and its most significant aftershocks ($M \geq 5$) are required. From about 150 aftershocks with magnitudes in the range of $4 \leq M < 5$, focal mechanisms of four events only have been reported by GCMT. So, we ignored them because of incompleteness issues and just considered events with magnitudes equal to or greater than 5, since either a slip model or a focal mechanism was available for all later. This information is available because many geoscientific studies have already been conducted due to the large magnitude of the Azgeleh main shock and its long-lasting aftershock sequence. Among them, many studies investigated the coseismic or the combined coseismic and post-seismic deformation and derived corresponding slip models using InSAR-based techniques (Barnhart et al. 2018; Feng et al. 2018; Wang & Burgmann 2018; Vajedian et al. 2018; Chen et al. 2018; Ding et al. 2018; Kuang et al. 2019; Yang et al. 2019; Nissen et al. 2019; Gombert et al. 2019; Fathian et al. 2021). For the Azgeleh main shock, Barnhart et al. (2018) proposed a model of a ramp-flat fault with a lower dip of 1–5° for the updip afterslip. The same authors also inverted coseismic InSAR displacements of the 11 January 2018 earthquake sequence (No. 4–8 in Table 1) that occurred ~80 km south of the Azgeleh earthquake epicentre. They find that the obtained moment magnitude from inversion for the related five moderate magnitude events ($M_w$ 5.0–5.5) is substantially larger than the moment magnitude of the largest event in the sequence ($M_w$ 5.8 versus $M_w$ 5.5). Therefore, the distributed slip inversions may represent the cumulative slip of these five events. Fathian et al. (2021) studied the coseismic permanent displacements of the Azgeleh main shock and following two large events (Tazehabad and Sarpol-e Zahab events) by using two-pass differential SAR interferometry (DInSAR) and Small Baseline Subset (SBAS) methods. They used non-linear and linear optimization algorithms to derive the source geometry and slip distribution over the fault planes.

For the Azgeleh main shock and its two largest aftershocks, we used the slip models derived by Fathian et al. (2021). They proposed a listric fault geometry for the main shock, with a shallow angle plane (eastern plane) dipping $\sim$3°, a ramp dipping $\sim$16° (central plane), and a steeper ramp dipping 25° (western plane). Their non-linear and linear inversion of the downsampled data showed an east–west trending fault with a dip to the north for the Tazehabad event. Fathian et al. (2021) also showed that the best-fitting inversion on the coseismic surface deformations of the Sarpol-e Zahab event agrees well with a single NE trending fault plane dipping to the southeast (Fig. S4). The 25 November 2018 Sarpol-e Zahab earthquake was shortly followed by two aftershocks with similar epicenters (No. 14 and 15 in Table 1, Fig. S4). Thus, the retrieved LOS displacement map includes the displacements of these two aftershocks. Therefore, we used the slip model obtained by Fathian et al. (2021) to represent the combined slip of the Sarpol-e Zahab and its two big aftershocks (No. 13, 14 and 15 in Table 1).

For earthquakes for which no slip models have been published (No. 2–8, 9, 10, 11, 16, 17 and 18 in Table 1), we used the available focal mechanisms from GCMT catalogue in combination with empirical scaling relationships to obtain the fault dimensions and mean slip values (Hanks & Kanamori 1979; Wells & Coppersmith 1994). For the 11 January 2018 sequence (No. 4–8 in Table 1), we used the cumulative moment obtained by Barnhart et al. (2018) [same as derived by Nissen et al. (2019)] and considered the focal plane with eastward slip to be the rupture plane. In general, for events with no available slip model, we considered the focal mechanism plane as the rupture plane that best matches the predominant tectonics of the study area. The 11 December 2017 ($M_w$ 5.5) event (No. 1 in Table 1) involves NW–SE left-lateral or NE–SW right-lateral strike-slip, and according to Nissen et al. (2019), the left-lateral nodal plane is conspicuously parallel to the structural fabric of the overlying fold axes. The 22 July 2018 ($M_w$ 5.8) event (No. 11 in Table 1) occurred near the Tazehabad earthquake. Thus, similar to the Tazehabad event, we considered the E–W trending plane as its fault rupture plane. For all other events, the nodal planes with NW–SE trending dip to the east are considered as the fault planes (Fig. S5).

3 STRESS METRICS AND DISTANCE MODEL

For each of the source models described in the last section and listed in Table 1, we calculated the static stress tensor on the grid points of our target region. For this purpose, we used the PSGRN+PSCMP tool of Wang et al. (2006) to determine the coseismic stress changes in a layered half-space. In our analysis, the layering is based on the CRUST 2.0 velocity model (Bassin 2000). However, we also tested a local velocity structure derived specifically for the region (Fathian et al. 2021), and found that our results are stable.

We resolved the stress tensors on a 3-D regular grid with either 5 or 2 km spacing in the horizontal and depth directions covering the region up to 75 km away from the main shock rupture and depths up to 20 km. At each grid point, we calculated the distance-slip value (R) proposed by Mignan & Broccardo (2019) and the following five stress scalars: The receiver-dependent metrics (i) $\Delta$CSF on master fault orientation (MAS), (ii) $\Delta$CSF on OOPs and (iii) $\Delta$CSF assuming fault variability (VM), and the receiver-independent stress scalars (iv) maximum shear (MS) and (v) von-Mises stress (VMS).
For the three metrics based on ΔCFS, we used the constant apparent friction model of Cocco & Rice (2002). According to this model, the Coulomb stress changes can be written as ΔCFS = Δτ − μΔσ, where τ refers to the shear stress in slip direction, σ is the normal stress on the fault plane, which is positive for compression and μ̄ is the effective friction coefficient (King et al. 1994). In our application, we set μ̄ = 0.4 but we also test it for values of 0.2 and 0.6.

For MAS, ΔCFS was calculated for a receiver mechanism identical to the main shock mechanism, which agrees with the predominant tectonic of the study area (i.e. the Zagros continental collision zone). For OOP, ΔCFS is calculated on the planes that maximize, for the given friction coefficient, the total Coulomb stress which consists of the sum of the pre-existing regional stress and the main shock-induced stress change. So, a background stress field with its orientation and strength needs to be defined. Here, we set the orientation of the principal stress components so that the main shock was optimally oriented to the background stress field. Furthermore, we use a differential stress of 3 MPa (σ1 = 1, σ2 = 0 and σ3 = −2 MPa), which agrees with the average stress drop of interplate earthquakes (Allmann & Shearer2018). In reality, the background stresses might be larger. However, in the case of a large value of the background stress, the OOP planes would be determined by the background stress field and the result of OOP would be identical to MAS. Thus, to analyse the difference to MAS, we chose a rather small value. For VM, in each grid node 1500 planes were randomly selected from Gaussian distributions centered around the main shock mechanism with standard deviations of 30° for the strike, dip, and rake. Assuming that earthquakes can only occur on receivers with positive stress changes, the VM value was calculated by summing only the positive stresses divided by the total number of selected planes. The maximum change in shear stress (MS) is the scalar maximum of the deviatoric stress tensor. The VMS is a scaled version of the second invariant of the deviatoric stress-change tensor related to the main shock-induced stress changes, which quantifies the yield limit of material under load (DeVries et al. 2018).

In addition to the stress scalars, Mignan & Broccardo (2019) introduced a distance-slip model (R), which does not need any stress calculation. The R-model was derived by a logistic regression based on the average slip (d) and the minimum distance (r) between the fault and grid node. According to this model, the probability of aftershock occurrence at a location is proportional to

\[ P(r, d) = \frac{1}{1 + e^{-[\beta_0 + \beta_1 \log(r) + \beta_2 \log(d)]}} \]

with parameters β0 = 10.18 ± 0.07, β1 = −2.32 ± 0.02 and β2 = 1.16 ± 0.01. We used these parameter values also in our study.

4 TEST METHODS

To assess the ability of each stress metric and the distance-slip model to explain aftershock locations, we need a classification model which transforms each element xi of the input into a binary prediction y'i: positive (aftershocks) or negative (no aftershocks). This can be done by scoring classifiers that produce a real-valued prediction score f(xi) for each element. It assigns a positive prediction (y'i = 1) when the score exceeds some threshold τ, or negative prediction (y'i = 0) otherwise (Fawcett 2006; Cao et al. 2020). Given a prediction and the observation of whether or not aftershocks occurred in the given cell, there are four possible outcomes: (1) true positive (TP: the prediction and observation are both positive); (2) false positive (FP: the prediction is positive, but the observation is negative); (3) false negative (FN: the prediction is negative, but the observation is positive) or (4) true negative (TN: both the prediction and observation are negative). By comparing a list of predictions with a set of observations, a 2 × 2 confusion matrix (also called a contingency table) can be constructed that summarizes classifier performance at a particular threshold τ and forms the basis for many common metrics (Table 2).

The most common ways for assessing and comparing binary classifier performance are the receiver operating characteristic (ROC) curve (Kleinbaum & Klein 1994) and the precision-recall (PR) curve (Raghavan et al. 1989). In previous analyses of aftershocks spatial forecasts, the ROC curves were used to evaluate the forecasting capability of different methods such as primary physics-based models including ΔCFS, neural network predictions, and the R-model (Meade et al. 2017; DeVries et al. 2018; Mignan & Broccardo 2019; Sharma et al. 2020). The numbers of TP, FP, FN and TN are counted for different detection thresholds, and the true positive rate [TPR = TP/(TP + FN)] and false positive rate [FPR = FP/(TN + FP)] are calculated. By plotting calculated TPR values against the FPR values, the ROC curves are created. The area under the ROC curve (AUC) is used to evaluate the prediction method and to compare different models. A random classification of binary data would result in a straight-line ROC curve with a slope of 1.0, corresponding to AUC = 0.5. Thus a model with AUC > 0.5 is better than a random classifier, and a model with AUC < 0.5 can be rejected.

Despite the many advantages of the ROC curves over a statistic like the mis-classification rate (Meade et al. 2017), some studies show that the ROC analysis on imbalanced data can be misleading (Davis et al. 2005; Davis & Goadrich 2006; Jeni et al. 2013; Saito & Rehmsmeier 2015). In particular, Parsons (2020) discussed that the previous ROC tests for spatial aftershock forecasts suffer from data imbalance because most areas lack any event (many more negative than positive cases), inhibiting the resolving power of the ROC analysis. Recently Cao et al. (2020) addressed this problem and proposed as an alternative the MCC-F1 curve that combines two informative single-threshold metrics, Matthews correlation coefficient (MCC) and the F1 score

\[ F_1 = \frac{2TP}{2TP + FP + FN} \]

\[ MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \]

While most confusion matrix metrics such as the F1 score (eq. 2) range in the [0, 1] interval, the MCC score (eq. 3) ranges in the [−1, +1] interval. To transform the MCC also to the range [0, 1], Cao et al. (2020) rescaled the MCC using the unit-normalized value by replacing MCC by (MCC+1)/2.

The MCC-F1 curve plots the rescaled MCC against the F1 across the full range of possible thresholds. Unlike the ROC and PR curves, which combine metrics that only capture a single aspect of the confusion matrix, both MCC and F1 scores summarize the whole confusion matrix. Thus, MCC-F1 especially proves more informative when using imbalanced data sets. The MCC-F1 curve by visualizing the MCC and F1 score across the full range of possible thresholds enables comparing classifiers more comprehensively and reliably than the ROC and PR curves, better differentiating good and bad classifiers (Cao et al. 2020).
Table 2. Confusion matrix (contingency table). The frequency of every combination of predicted class \( y'_i \) and actual class \( y_j \) are quantified by the four cells of the confusion matrix for a binary classifier.

<table>
<thead>
<tr>
<th>Actual positive ((y_j = 1))</th>
<th>Actual negative ((y_j = 0))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted positive ((y'_i = 1))</td>
<td>True positives (TP)</td>
</tr>
<tr>
<td>Predicted negative ((y'_i = 0))</td>
<td>False negatives (FN)</td>
</tr>
</tbody>
</table>

For the MCC-F1 curve, a random classifier is related to a unit-normalized MCC value of 0.5. Thus, any reasonable classifier needs to be above the horizontal line at 0.5. The points of perfect and worst performance are \((1,1)\) and \((0,0)\), respectively, which predict all data elements correctly or incorrectly (Cao et al. 2020). To quantify the quality of MCC-F1 curves, Cao et al. (2020) introduced a metric based on the distance between the MCC-F1 curve and the point of perfect performance \((1,1)\), which is given by

\[
\text{MCC} - \text{F1 metric} = 1 - \frac{D}{\sqrt{2}}, \quad (4)
\]

where \(D\) is the average distance to point \((1,1)\) along the curve and \(\sqrt{2}\) is the distance between the points of worst \((0,0)\) and perfect \((1,1)\) performance [see details for the calculation in Cao et al. (2020)].

The worst value for the MCC-F1 score is 0, and the best value is 1. Therefore, better classifiers have larger MCC-F1 metrics, and their curves are closer to the point of perfect performance \((1,1)\). The MCC-F1 value for random classifiers is not generally defined. While MCC equals 0.5, the F1 score varies for random classifiers, as it includes precision which depends on the class balance of the test set. Instead, the AUC metric should be used to judge whether a classifier is better than random, which yields 0.5 regardless of the class balance (Fawcett 2006).

We use both ROC and MCC-F1 analysis to quantify the accuracy of the aftershock forecasts. The inhomogeneity of the earthquake catalog due to varying completeness over time and space, associated location uncertainties, and the occurrence of background activity, which is not related to the main shock, can bias the classifiers’ results. Note that we did not explicitly model the short-term temporal incompleteness in the first days after major events because it is expected to play not a crucial role in our analysis. If the incompleteness similarly affects the whole region, the total number is reduced but not the spatial distribution. However, to test the potential effect of these issues for the Kermanshah sequence, we calculated the AUC and MCC-F1 metrics for different aftershock periods and magnitude cut-offs. Furthermore, we explore the potential dependence of the results on the grid resolution (10, 5 and 2 km).

5 RESULTS

5.1 Forecasts based solely on the main shock source information

We first considered only the main shock source information to calculate the stress- and the distance-slip-based forecasts for the \(5 \times 5 \times 5\) km gridded target volume. In particular, we performed a detailed analysis of the ROC and MCC-F1 values for the aftershock data that occurred during the whole study period (12.11.2017–04.07.2019). The resulting ROC and MCC-F1 curves are plotted in Figs 3(a) and (a’), respectively. Based on the ROC analysis, the best performing metrics are the maximum shear and von-Mises scalars (AUC = 0.758). The AUC values for the VM model with ΔCFS calculated on distributed planes and OOP model with ΔCFS calculated on OOPs are slightly lower but also perform well (AUC = 0.752 and 0.745, respectively), similarly the R model (AUC = 0.746). In contrast, the classic Coulomb stress change MAS model (AUC = 0.609) is worse but still better than a random classification (Fig. 3a). On the other hand, the results are different for the MCC-F1 analysis. In this case, the VM (metric = 0.398) outperforms other stress metrics, and the R model (metric = 0.410) has the best performance (Fig. 3a’).

In the second step, we repeated the same analysis for a finer spatial grid with a spacing of 2 km in horizontal and depth directions to test the dependence of the ROC and MCC-F1 results on the spatial resolution. The corresponding results for the higher-resolution \(2 \times 2 \times 2\) km grid are presented in Figs 3(b) and (b’). For the ROC analysis, the performance of all stress metrics and the R model increases, and OOP with the most increase (AUC = 0.799) performs similarly to MS and VMS. The results for the MCC-F1 analysis are opposite, with decreasing values for all forecast models (Fig. 3b’). However, for the higher-resolution grid, the R model also shows the best MCC-F1 result. We repeated the calculation for a coarser \(10 \times 10 \times 10\) km grid and found a systematic dependence of the ROC and MCC-F1 results on the grid size. While all AUC values of the ROC analysis increase by increasing the grid resolution, MCC-F1 results have the opposite trend. Furthermore, the results show that the difference between the OOP, VM, MS and VMS becomes insignificant in both ROC and MCC-F1 analysis for higher-resolution grids (Fig. S6 and Table S1). However, in all cases, the MAS model is the worst for both test metrics, while the R model always has the best MCC-F1 results, but it is always worse than the best stress scalars in the ROC analysis.

In the next step, we checked the dependence of the AUC and MCC-F1 values on the magnitude cut-off (Mcut). For grid spacing of 2 km, we compared the result for the complete catalog (Mcut = 1) with the corresponding results for higher cut-offs 1.5, 2.5, 3.5 and 4.5 (Fig. S7). Figs 3(c) and (c’) show the ROC and MCC-F1 results for Mcut = 2.5, corresponding to the estimated completeness magnitude \(M_c\). In comparison to Figs 3(b) and (b’), it can be seen that the ROC and MCC-F1 results decrease for all models, except the AUC value for OOP. According to the AUC values, the OOP model also remains the best model for higher-magnitude cut-offs. In contrast, the R model loses its leading position for MCC-F1, and all stress metrics besides MAS become slightly better than R for higher Mcut values.

The results might be affected by background activity not related to the main shock. The percentage of background activity is expected to increase with increasing time after the main shock because of the Omori decay of the aftershock rate. To analyse the potential impact of background activity, we thus repeated our calculation for shorter target periods \((T)\); in particular, we analysed the ROC and MCC-F1 results for \(T = 6, 12\) and 20 months after the Azgeleh main shock (Figs 3d and d’ show the results for the first year aftershocks). We find that the ROC results are almost independent of \(T\), while the MCC-F1 metric slightly increases for all models with increasing \(T\) (Fig. S8 and Table S3). It seems that MCC-F1 metric is dependent on the number of aftershocks, which naturally increases with \(T\).
Figure 3. ROC (left-hand column) and MCC-F1 (right-hand column) results using only the Azerleah main shock source information for different grid spacings (L), magnitude cut-offs (Mcut) and aftershock periods (T). (a, a’) L = 5 km, Mcut = 1 and T = 600 d. (b, b’) L = 2 km, Mcut = 1 and T = 600 d. (c, c’) L = 2 km, Mcut = 2.5 and T = 600 d. (d, d’) L = 2 km, Mcut = 1.0 and T = 1 yr. In all panels, the dashed line represents the result for a random classifier.

In order to test the robustness of the results to other choices of friction values, we repeated the analysis for different values. For the receiver-dependent stress metrics (MAS, OOP and VM), we considered the additional friction values 0.2 and 0.6 using a spatial resolution of 5 km. Our results show that both ROC and MCC-F1 results slightly improve for the smaller friction value (0.2), while their performance slightly decreases for the larger value (0.6). However, the relative changes are minor, not exceeding 1.1 per cent.
relative to the results for our standard choice of 0.4 in all cases (Table S4).

Ideally, the forecasts would be based on values representing the average (or integrated) value within the forecast cells. However, due to computational costs, we only used the value at the centre point as an estimator of the average, which works well if stress (or distance metrics) varies approximately linearly within the cell. However, differences are expected close to the earthquake ruptures where stress varies strongly non-linear. To check our approximation, we run a test using the $2 \times 2 \times 2$ km gridded volume. We compared the forecast results based on the center point with those based on the averages of 8 points separated with 1 km spacing within each cell. We find that the relative change of all AUC and MCC-F1 values is minor (<2 per cent for MAS and <1 per cent for all other models), and the worst and best models remain the same. Thus, we conclude that approximating the average stress by the stress at the centre point works reasonably well.

5.2 Forecasts based on multiple source information

Before, the forecast models relied only on the source information of the main shock. We repeated the calculations and re-analysed the results considering the additional source information of the largest of the main shock. We repeated the calculations and re-analysed the forecasts for multiple sources; instead, ROC generally worsens. On the other hand, the R model expects the relative change of all AUC and MCC-F1 values is increased values in the near as well as the far-field. The spatial patterns are very similar, explaining why their ROC and MCC-F1 results are similar. On the other hand, the R model expects a higher aftershock density near the rupture with monotonously decreasing values with increasing distance (Fig. 4f). By comparing these maps with the patterns of stress scalars due to just the main shock (Fig. S9), the secondary earthquake sources are found to increase the stress scalars, particularly close to the $M \geq 5$ epicentres, which is obvious especially for the seismicity pattern in the southern part of the study area.

The AUC and MCC-F1 results for the forecasts based on multiple sources information are shown in Fig. 5 for a grid spacing of $5 \times 5 \times 5$ km. The corresponding results for $2 \times 2 \times 2$ km are shown in Fig. S10. For comparison, the solid line in each panel refers to the forecast based solely on the main shock source information. In contrast, the forecast results based on multiple sources are indicated by dashed lines for equal weights ($w_i = 1$) and dotted lines for Omori-weights (eq. 6).

For MAS and OOP, the results are not improved by accounting for multiple sources; instead, ROC generally worsens. On the other hand, the other stress scalars (VM, MS and VMS) and distance-slip model (R) significantly improve by considering additional information from preceding large aftershocks. The forecasts for the first period are the same because only the main shock source information is available. Also, the information of the first $M \geq 5$ aftershock does not change the forecasts significantly. However, after the second and third $M \geq 5$ aftershocks, the forecasts using multiple sources start to significantly differ and outperform the forecast based solely on the main shock. Furthermore, the forecasts based on Omori-weighted multiple sources outperform the uniformly weighted forecasts systematically in the case of the best performing stress scalars VM, MS and VMS for both AUC and MCC-F1 metrics.

Fig. 6 compares the forecast results based on Omori-weighted multiple sources in the same panel, facilitating the comparison of the best stress scalars and the R model. It shows that the VM, MS and VMS forecasts outperform the R model concerning AUC, while they perform overall similarly for MCC-F1.

Finally, we also tested the case that the forecasts are based solely on the source information of the latest $M \geq 5$, that is the weights are $w_i = \delta_{ik}$ with Kronecker delta $\delta_{ik} = 1$ if $i = k$ and 0 else. The results show that this model generally outperforms the uniformly weighted model, but it is worse than the Omori-weighted model in most periods. In general, the overall best forecasts are found for the Omori-weighted MS and VMS forecasts (Figs 5 and 6).

6 DISCUSSION

Accurate spatial aftershock forecasts are important for planning rescue activities and risk mitigation. Thus, different forecast models...
Figure 4. Superimposed forecast maps using the main shock and large aftershocks \((M \geq 5)\) during 2017–2019 Kermanshah sequence at the hypocentral depth of 9.0 km. (a–f) Maps showing the values of MAS, OOP, VM, MS, VMS and R, respectively. Black circles indicate areas where one or more aftershocks occurred in the ±1.0 km depth interval during the analysed sequence and their distance in the dense area give an imagination of the grid size. The big black star refers to the Azgeleh main shock epicentre, and the small yellow stars refer to the large events \((\geq 5)\) epicentres.

should be rigorously tested before their potential application in real cases. However, test results might depend on the test metric, such as AUC and MCC-F1. Aftershock distributions are spatially imbalanced, with almost always many more negative cases (places with no earthquakes). Consequently, the ratio between TNs and TPs is out of balance. As Parsons (2020) discussed, the ROC test specifically suffers from data imbalance of aftershock data. The ROC tests will score particularly poorly for the classical Coulomb model MAS, predicting areas with negative stress changes where aftershocks should be unlikely. The reason is that ROC does not as strongly penalize FPs. Based on Davis & Goadrich (2006), in the case of large imbalance, ROC curves present an overly optimistic view of a classifier’s performance. For aftershock forecasts with a large number of actual negatives, the data is negatively imbalanced. In this case, large relative changes in false positives only lead to minor changes in the FPR value. Our analysis for the 2017–2019 Kermanshah earthquake sequence shows that the AUC values of the ROC curves are close to one, that is the best value, and thus the outcome is very optimistic. The problem of data imbalance becomes obvious by our tests with different grid resolutions. The AUC values of all models systematically increase with decreasing grid-spacing (Fig. 3). Increasing the grid resolution increases the number of grid cells, while the total number of aftershocks bounds the TP value. Thus, the data imbalance increases because of the increased FPs, and the observed increase of AUC is unreliable. However, the ROC test may still be useful for comparing models.

Cao et al. (2020) showed that MCC-F1 analysis is more informative than ROC analysis on a negatively imbalanced data set. In our data set with a large number of actual negatives, the results of the MCC-F1 metric seem thus more realistic, which is also reflected by decreased MCC-F1 results decrease for increased data imbalance in the case of higher grid-resolutions. Furthermore, in contrast to the ROC analysis, the MCC-F1 metric can be used to recognize the best threshold for each model in addition to model comparisons. Similarly, the tendency of our test results for increasing cut-off magnitudes and shortened aftershock periods can be understood. In both cases, the number of target events decreases, leading to more FPs and fewer TPs. Therefore, the performance power of a reasonable classifier should decrease. In contrast to AUC, the MCC-F1 metric shows this tendency, justifying it as a classifier for binary aftershock forecasts.

The model ranking for both AUC and MCC-F1 yields that MAS provides the worst results. The underlying reason for the poor performance of classical Coulomb stress changes compared to the other stress scalars is still unclear and needs more investigation. However, it seems that the occurrence of aftershocks in stress shadows
Figure 5. AUC (a) and MCC-F1 (b) results for the different stress metrics and R model for 5 km grid-spacing. In each panel, the solid line refers to the forecasts based solely on the main shock source information. In contrast, the forecasts based on multiple sources are indicated by dashed lines in the case of uniform weights and dotted lines for Omori-type weights.

(FNs) reduces the predictive power of the MAS model. The reasons for seismicity in those shadow zones can be manifold, including dynamic stress triggering (Hill & Prejean 2015) and post-seismic stress changes related to fault afterslip, viscoelastic relaxation (Pollitz 1992), or aftershocks (secondary triggering). Furthermore, uncertainties of the slip model, aftershock locations and receiver mechanisms can also explain aftershocks within apparent stress shadows (Hainzl et al. 2010).

Our analysis confirms that stress changes by large aftershocks can play an important role in aftershock forecasts. Inclusion of large aftershocks ($M \geq 5$) in our analysis improved both AUC and MCC-F1 classifier results for stress scalars and the R model (Fig. 5). Simple superimposing the model results from the main shock and large aftershocks (uniform weights) improves the forecasting results. However, they improve more strongly for Omori-weighted forecasts, especially for the best stress scalars VM, MS and VMS. The latter is reasonable because we know that the effect of stress changes is transient, with an Omori-type decay (Dieterich 1994). In principle, it would be desirable to include the stress effects of all observed earthquakes in stress-based forecasts. However, unfortunately, no reliable slip models are usually available for small to moderate events. One possible solution is to assume simplified isotropic changes, which decay with distance, for those events (Cattania et al. 2015). In this way, detailed triggering forecasts for reliable source information could be combined with forecasts for smaller events solely based on the catalogue information of occurrence time, hypocentre and magnitude. Indeed, a natural extension of our current work would be the implementation of the stress metrics and R model in the
spatiotemporal epidemic type aftershock sequence (ETAS) model (Ogata 1998; Bach & Hainzl 2012), in which the triggering effect of each earthquake also decays with time according to the Omori law.

7 CONCLUSIONS

Aftershock forecasts are an important ingredient for time-dependent seismic hazard assessments. Their occurrence is mainly related to the static stress changes of the main shock, although other effects such as dynamic or post-seismic processes and alterations of fault strengths can also play an important role. Recent analyses of binary forecasts of the aftershock area revealed that the classical approach of calculating ΔCFS is not the best and outperformed by receiver-independent stress scalars (Meade et al. 2017; Sharma et al. 2020). However, these results were questioned because of the applied ROC analysis for testing, which might suffer from data imbalance (Parsons 2020).

Our detailed analysis of the complex Kermanshah sequence confirms that the static stress-triggering hypothesis leads to informative forecasts, but with worst results for ΔCFS resolved on a fixed receiver mechanism. In contrast, we obtained the best results for the VM model accounting for variable mechanisms as well as the receiver-independent MS and VMS stress scalars. Furthermore, the R model, which relies only on the mean slip and the distance to the rupture, is among the best models, as observed in previous studies (Sharma et al. 2020). These results are confirmed by the more appropriate test metric MCC-F1 proposed by Cao et al. (2020).

In general, different metrics may work better in different sequences. Thus, an ensemble approach with multiple metrics could lead to more robust forecasts. Whether the MCC-F1 values are useful ensemble weights must be tested in future studies.

In addition to previous studies, we analysed the impact of large aftershocks on the test results. For that purpose, we created weighted forecasts for multiple sources, with either uniform or Omori-type weights. We find that the classical ΔCFS approach remains the worst and is not improved. However, the Omori-type weighted multiple-source forecasts significantly outperform all other models for the most informative stress scalars, indicating the importance of considering secondary stress changes for aftershock forecasting. Hybrid models using the improved stress scalars in the ETAS framework are thus a promising approach to improve aftershock forecasts in the future.

ACKNOWLEDGMENTS

We thank Dr Margarita Segou (Associate Editor), Dr Andrew Michael and anonymous reviewer for their thoughtful comments and suggestions that are valuable and very helpful for revising and improving this paper. This work was supported by the International Institute of Earthquake Engineering and Seismology (IIIES) in the framework of the PERSIA (Probabilities of Earthquake Ruptures in Iran) project and the DFG - Research Training Group NaRiskChange. We acknowledge the Iranian Seismological Center (IRSC) for providing the earthquake sequence data. S.H. was supported by the European Unions Horizon 2020 research and innovation program under Grant Agreement Number 821115, real-time earthquake risk reduction for a resilient Europe (RISE).

Author Contribution: This work is a part of the PhD thesis of BMA. Investigation and research were performed by all authors. Specifically, the ideas and research goals of this manuscript were formulated by BMA in the leadership of co-authors. BMA wrote the codes, processed statistical analysis and obtained the first results. He wrote the original draft under the supervision of all co-authors. All authors contributed significantly in many ways, for example by critical review and revisions.

DATA AVAILABILITY

Slip data obtained by Fathian et al. (2021) for the large events are freely available on https://www.sciencedirect.com/science/article/abs/pii/S0034425720305976 and data for aftershocks are from Iranian Seismological Center (IRSC) event catalogue (http://irsc.ut.ac.ir/bulletin.php).

REFERENCES


SUPPLEMENTARY INFORMATION

Supplementary data are available at GJI online.

Figure S1. Focal mechanism of Azgeleh main shock (black beach ball) and spatial and depth distribution of seismicity until the Taizehabad earthquake (25.08.2018). The top plot shows the magnitude versus time of events that occurred within a distance of 75 km to the Azgeleh main shock rupture with a depth of less than 20 km during the 2017–2019 Kermanshah sequence. Colourful circles show the events that occurred between the Azgeleh and Taizehabad main shocks (12.11.2017–25.08.2018), and white circles are for
the events that occurred after the Tazehabad event. The other symbols and lines are explained in the caption of Fig. 1 of the main manuscript.

**Figure S2.** Focal mechanism of Tazehabad main shock (black beach ball) and spatial and depth distribution of seismicity until the Sarpol-e Zahab earthquake (25.11.2018). The top plot is the same as Fig. S1, and colourful circles show the events that occurred between the Tazehabad and Sarpol-e Zahab main shocks (25.08.2018–25.11.2018), black circles show the events before the Tazehabad events and white circles are for the events that occurred after the Sarpol-e Zahab event. The other symbols and lines are explained in the caption of Fig. 1 of the main manuscript.

**Figure S3.** Focal mechanism of Sarpol-e Zahab main shock (black beach ball) and spatial and depth distribution of seismicity until the end of the sequence (04.07.2019). Top plot is the same as Fig. S1 and colourful circles show the events that occurred between the Sarpol-e Zahab main shock and the end of this sequence (25.11.2018–04.07.2019), black circles show the events before the Sarpol-e Zahab events. The other symbols and lines are explained in the caption of Fig. 1 of the main manuscript.

**Figure S4.** Focal mechanism of the Azgeleh, Tazehabad, and Sarpol-e Zahab main shocks (black beach balls) and 15 large aftershocks with magnitude equal to or larger than 5 (blue beach balls). The other symbols and lines are explained in the caption of Fig. 1 of the main manuscript.

**Figure S5.** Slip model of the used events in this study.

**Figure S6.** ROC (top row) and MCC-F1 (bottom row) results of the different stress scalars and R model for different grid spacings: (a, a’) $10 \times 10 \times 10$ km; (b, b’) $5 \times 5 \times 5$ km and (c, c’) $2 \times 2 \times 2$ km. The dashed lines refer to the results for random classifiers.

**Figure S7.** ROC and MCC-F1 results of the different stress scalars and R model for different magnitude cut-offs (2 × 2 × 2 km grid spacing): (a, a’) $M_{\text{cut}} = 1$; (b, b’) $M_{\text{cut}} = 1.5$; (c, c’) $M_{\text{cut}} = 2.5$; (d, d’) $M_{\text{cut}} = 3.5$ and (e, e’) $M_{\text{cut}} = 4.5$. The dashed lines refer to the result for random classifiers.

**Figure S8.** ROC (top row) MCC-F1 (bottom row) results of the different stress scalars and R model for different aftershock periods $T$ and $2 \times 2 \times 2$ km grid spacing: (a, a’) $T = 20$ months; (b, b’) $T = 12$ months; (c, c’) $T = 6$ months. The dashed lines refer to the result for random classifiers.

**Figure S9.** The same as manuscript Fig. 6, but for maps solely based on the Azgeleh main shock.

**Figure S10.** The same as manuscript Fig. 7, but for 2 km grid-spacing.

**Table S1.** Dependence of the AUC and MCC-F1 values on the grid spacing.

**Table S2.** Dependence of the AUC and MCC-F1 values on the magnitude cut-off in the case of a grid with 2 km spacing.

**Table S3.** Dependence of the AUC and MCC-F1 values on the aftershock period (2 km grid-spacing).

**Table S4.** Dependence of the AUC and MCC-F1 values on the friction value.

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the paper.