1 A Wavelet-Based Approach to Streamflow Event Identification and Modeled Timing Error
2 Evaluation

4 Erin Towler[1,*] and James L. McCreight[1,2]
5 [1] National Center for Atmospheric Research (NCAR), P.O. Box 3000, Boulder, CO 80307
6 [*] Corresponding author: towler@ucar.edu, https://orcid.org/0000-0002-1784-1346
7 [2] orcidid: 0000-0001-6018-425X

8 **Abstract**

10   Streamflow timing errors (in the units of time) are rarely explicitly evaluated, but are

11 useful for model evaluation and development. Wavelet-based approaches have been shown to

12 reliably quantify timing errors in streamflow simulations, but have not been applied in a

13 systematic way that is suitable for model evaluation. This paper provides a step-by-step

14 methodology that objectively identifies events, and then estimates timing errors for those events,

15 in a way that can be applied to large-sample, high-resolution predictions. Step 1 applies the

16 wavelet transform to the observations, and uses statistical significance to identify observed

17 events. Step 2 utilizes the cross-wavelet transform to calculate the timing errors for the events

18 identified in Step 1; this includes the diagnostic of model event "hits", and timing errors are only

19 assessed for hits. The methodology is illustrated using real and simulated stream discharge data

20 from several locations to highlight key method features. The method groups event timing errors

21 by dominant timescales, which can be used to identify the potential processes contributing to the

22 timing errors and the associated model development needs. For instance, timing errors that are

23 associated with the diurnal melt cycle are identified. The method is also useful for documenting

24 and evaluating model performance in terms of defined standards. This is illustrated by showing

25 version-over-version performance of the National Water Model (NWM) in terms of timing

26 errors.

28 **1. Introduction**

30   Common verification metrics used to evaluate streamflow simulations are typically

31   aggregated measures of model performance, e.g., the Nash Sutcliffe Efficiency (NSE) and the

32   related root mean square error (RMSE). Although typically used to assess errors in amplitude,

33   these statistical metrics include contributions from errors in both amplitude and timing (Ehret

34   and Zehe 2011), making them difficult to use for diagnostic model evaluation (Gupta et al.

35   2008). Furthermore, common verification metrics are calculated using the entire time series,

36   whereas timing errors require comparing localized features or events in the data. This paper

37   focuses explicitly on event timing error estimation, which is not routinely evaluated, despite its

38   potential benefit for model diagnostics (Gupta et al. 2008) and practical forecast guidance (Liu et

39   al. 2011).

40   The fundamental challenge with evaluating timing errors is identifying what constitutes

41   as an "event" in the two time series being compared. Identifying events is typically subjective,

42   time consuming, and not practical for large-sample hydrological applications (Gupta et al. 2014).

43   A variety of baseflow separation methods, ranging from physically-based to empirical, have been

44   developed to identify hydrologic events (see Mei and Anagnostou 2015 for a summary), though

45   many of these approaches require some manual inspection of the hydrographs. Merz et al. (2006)

46   put forth an automated approach, but it requires a calibrated hydrologic model, which is a

47   limitation in data poor regions. Koskelo et al. (2012) developed a simple, empirical approach that

48   only requires rainfall and runoff time series, but is limited to small watersheds and daily data.

49   Mei and Anagnostou (2015) introduce an automated physically-based approach, which is

50   demonstrated for hourly data, though one caveat is that basin events need to have a clearly

51   detectable recession period. Additional methods have focused on identifying flooding events

52   using peak-over-threshold methods. The thresholds used for such analyses are often either based

53   on historical percentiles (e.g., the 95th percentile) or on local impact levels (river stage), such as

54   the National Weather Service (NWS) flood categories (NOAA National Weather Service, 2012).

55   Timing error metrics are often calculated from the peaks of these identified events. For example,

56   the Peak Time Error, or its derivative the Mean Absolute Peak Time Error, requires matching

57   observed and simulated event peaks, and calculating their offset (Ehret and Zehe 2011). While

58   this may be straightforward visually, it can be difficult to automate; some of the reasons for this

59   are discussed below.

60   Difficulties arise using thresholds for event identification. For example, exceedances can

61   cluster if a hydrograph vacillates above and below a threshold, begging the question: Is it one or

62   multiple events? Which peak should be used for the assessment? In the statistics of extremes,

63   declustering approaches can be applied to extract independent peaks (e.g., Coles 2001), but this

64   reductionist approach may miss relevant features. For instance, if background flows are elevated

65   for a longer period of time before and after the occurrence of these "events", the threshold-based

66   analysis identifies features of the flow separately from the primary hydrologic process

67   responsible for the event. If one focuses just on peak timing differences in this example, that

68   timing error may only apply to some small fraction of the total flow of the larger event which

69   happens mainly below the threshold. Further, for overall model diagnosis that focuses on model

70   performance for all events, not just flood events, variable thresholds would be needed to account

71   for different kinds of events (e.g., a daily melt event versus a convective precipitation event).

72   Using a threshold-approach to identify events and timing error assessment, Ehret and

73   Zehe (2011) develop an intuitive assessment of hydrograph similarity, the Series Distance. This

74   algorithm is later improved upon by Siebert et al. (2016). The procedure matches observed and

75   simulated segments (rise or recession) of an event, and then calculates the amplitude and timing

76 errors, as well as the frequency of event agreement. The Series Distance requires smoothing the

77 time series, identifying an event threshold, and selecting a time range to consider two segments

78 matching.

79   Liu et al. (2011) developed a wavelet-based method for estimating model timing errors.

80 Although wavelets have been applied in many hydrologic applications such as model analysis

81 (e.g. Lane 2007; Weedon et al. 2015; Schaefli and Zehe 2009, Rathinasamy et al. 2014) and

82 post-processing (Bogner and Kalas 2007; Bogner and Pappenberger 2011), Liu et al. were the

83 first to use it for timing error estimation. Liu et al. (2011) apply a cross-wavelet transform

84 technique to streamflow time series for 11 headwater basins in Texas. Timing errors are

85 estimated for medium- to high- flow "events" that are determined a priori by threshold

86 exceedance. They use synthetic as well as real streamflow simulations to test the utility of the

87 approach. They show that the technique can reliably estimate timing errors, though they

88 conclude that it is less reliable for multi-peak or consecutive "events" (defined qualitatively).

89 ElSaadani and Krajewski (2017) followed the cross-wavelet approach used by Liu et al (2011) to

90 provide similar analysis and further investigate the effect of the choice of mother wavelet on the

91 timing error analysis. Ultimately, they recommended that in the situation of multiple, adjoining

92 flow peaks the improved time localization of the Paul wavelet might justify its poorer frequency

93 localization compared the Morlet wavelet.

94   Liu et al. (2011) provide a starting point for the work in this paper where we develop two

95 new bases for their method: 1) objective event identification for timing error evaluation and 2)

96 the use of observed events as the basis for the model timing error calculations. The latter is

97 important for "model benchmarking", i.e., the practice of evaluating models in terms of defined

98 standards (e.g., Luo, et al. 2012; Newman et al. 2017). Here, the use of observed events provides

99   a baseline by which to evaluate changes and to compare multiple versions or experimental

100   designs.

101   This paper provides a methodology for using wavelet analysis to quantify timing errors in

102   hydrologic simulations. Our contribution is a systematic approach that integrates 1) statistical

103   significance to identify events with 2) a basis for timing error calculations independent of model

104   simulations (i.e., benchmarking). We apply our method to timing error evaluation of high-

105   resolution streamflow prediction. The paper is organized as follows: Section 2 describes the

106   observational and simulated data used. Section 3 provides the detailed methodology of using

107   wavelets to identify events and estimate timing errors in a synthetic example. In Section 4, we

108   demonstrate the method using real and simulated streamflow data for several use cases, and then

109   illustrate the application of the method for version-over-version comparisons. Section 5 is the

110   discussion and conclusions, including how specific methodological choices may vary by

111   application.

112   **2. Data**

113
114   The application of the methodology is illustrated using real and simulated stream discharge

115   (streamflow, m3/s) data at three U.S. Geological Survey (USGS) stream gage locations in

116   different geographic regions: Onion Creek at US Highway 183, Austin, Texas, for the South

117   Central region (Onion Creek, TX; USGS site number 08159000), Taylor River at Taylor Park,

118   Colorado, for the Intermountain West (Taylor River, CO; USGS site number 09107000), and

119   Pemigewasset River at Woodstock, New Hampshire, for New England (Pemigewasset River,

120   NH; USGS site number 01075000). We use the USGS instantaneous observations averaged on

121   an hourly basis.

122   NOAA's National Water Model (NWM,

123   https://www.nco.ncep.noaa.gov/pmb/products/nwm/) is an operational model that produces

124   hydrologic analyses and forecasts over the continental United States (CONUS) and Hawaii (as of

125   version 2.0). The model is forced by downscaled atmospheric states and fluxes from NOAA's

126   operational weather models. Next, the NoahMP (Niu et al 2011) land surface model calculates

127   energy and water states and fluxes. Water fluxes propagate down the model chain through

128   overland and subsurface (soil and aquifer representations) water routing schemes to reach a

129   stream channel model. The NWM applies the three parameter Muskingum-Cunge river routing

130   scheme to a modified version of the NHD-Plus version 2 (McKay et al. 2012) river network

131   representation (Gochis et al 2020).

132   In this study, NWM simulations are taken from each version's retrospective runs

133   (https://docs.opendata.aws/nwm-archive/readme.html). These are continuous simulations (not

134   cycles) run for the period October 2010 to November 2016 and forced by the National Land Data

135   Assimilation System (NLDAS)-2 product as atmospheric conditions. The nudging data

136   assimilation was not applied in these runs. We use NWM discharge simulations from versions

137   V1.0, V1.1, and V1.2 (not all version may be publicly available).

138   The methodology developed in this paper is implemented in the R language and is made

139   publicly available, as detailed in the code availability section at the end of the manuscript.

140   **3. Methodology**
141
142   This section provides the description of the methodology using wavelets to identify events

143   and estimate timing errors. The steps can be seen in an accompanying flowchart (Figure 1) and

144   nomenclature table (Table 1), which defines key terms of the approach. To facilitate

145   understanding, the steps are illustrated by an application of the methodology to an observed time

146   series of an isolated peak in Onion Creek, TX, (Figure 2a) and the synthetic modeled time series

147   which is identical to the observation time series but shifted 5 hours in to the future (Figure 3a, note

148   the log scale).

149   *3.1. Step 1. Identify observed events*

150

151   The first step is to identify a set of observed events for which the timing error should be

152   calculated. We break this step into three sub-steps: 1a. Apply the wavelet transform to

153   observations, 1b. Determine all observed events using significance testing, and 1c. Sample

154   observed events to an event-set relevant to analysis.

155   3.1.1. Step 1a. Apply wavelet transform to observations

156   First, we apply the continuous wavelet transform (WT) to the observed time series. The

157   main steps and equations for the WT are provided here, though the reader is referred to Torrence

158   and Compo (1998) and Liu et al. (2011) for more details.

159   Before applying the WT, a mother wavelet needs to be selected. In Torrence and Compo

160   (1998), they discuss the key factors that should be considered when choosing the mother

161   wavelet. There are four main considerations, including (i) orthogonal or nonorthogonal, (ii)

162   complex or real, (iii) width, and (iv) shape. In this study, we follow Liu et al. (2011) in selecting

163   the nonorthogonal and complex Morlet wavelet:

164
$$\psi(n) = \pi^{-1/4} e^{iw_0 n} e^{-n^2/2},$$
165

166   where w_0 is the non-dimensional frequency, with a value of 6 (Torrence and Compo, 1998).

167   Once the mother wavelet is selected, the WT is applied to a time series x_n, where n goes

168   from n=0 to n=N-1, with a time step of δt. The WT is the convolution of the time series with the

169   mother wavelet that has been scaled and normalized:

$$170 \qquad W_n(s) = \sum_{n'=0}^{N-1} x_{n'} \psi^* \left[\frac{(n'-n)\delta t}{s} \right]$$

171

172   where n' is the localized time in [0, N-1], s is the scale parameter, and the asterix indicates the

173   complex conjugate of the wavelet function. The wavelet power is defined as $|W_n^2|$ which

174   represents the squared amplitude of an imaginary number when a complex wavelet is used as in

175   this study. We use the bias corrected wavelet power (Liu et al. 2007; Veleda et al. 2012), which

176   ensures power is comparable across timescales. We also identify a maximum timescale *a priori*

177   that corresponds to our application. We select 256 hours (~10 days), but this number could be

178   higher or lower for other applications and there are no real penalties for using too high a

179   maximum (lower than the annual cycle).

180   The wavelet transform (WT) expands the dimensionality of the original time series by

181   introducing the timescale (or period) dimension. Wavelet power is also a function of both time

182   and timescale (e.g. Torrence and Compo, 1998). This is illustrated in Figure 2: the streamflow

183   time series (panel a) is expanded into a 2-dimensional (2-D) wavelet power spectrum (panel b).

184   Wavelet analysis can detect localized signals in time series (Daubechies 1990), including

185   hydrologic time series, which are often irregular or aperiodic (i.e., events may be isolated and

186   don't regularly repeat) or non-stationary. We note that in many wavelet applications, timescale is

187   referred to as "period" and this axis is indeed the Fourier period in our plots. However, to

188   emphasize that our study is more focused on irregular events and less on periodic behavior of

189   time series, we use the term "timescale" to denote Fourier period (and not wavelet scale).

190   Because we are applying the WT to a finite time series, there are timescale-dependent

191   errors at the beginning and end times of the power spectrum, where the entirety of the wavelet at

192   each scale is not fully contained within the time series. This region of the WT is referred to as the

193   cone of influence or COI (Torrence and Compo, 1998). Figure 2b illustrates the COI as the

194   regions where the colors are muted; we ignore all results within the COI in this study.

195   We make several additional notes on the wavelet power and its representation in the

196   figures. The units of the wavelet power are those of the time series variance (m6/s2 for

197   streamflow) and it is natural to want to cast the power in a physical light or relate it to the time

198   series variance. Indeed, the power is often normalized by the time series variance when presented

199   graphically. However, it must be noted that the wavelet convolved with the time series frames

200   the resulting power in terms of itself at a given scale. Wavelet power is a (normalized) measure

201   of how well the wavelet and the time series match at a given time and scale. The power can only

202   be compared to other values of power resulting from a similarly constructed WT. There are

203   various transforms that can be applied to aid graphical interpretation of the power (log, variance

204   scaling), but the utility of these often depends on the nature of the individual time series

205   analyzed. For simplicity, we plot the raw bias-rectified wavelet power in this paper.

206   3.1.2. Step 1b. Determine all observed events using significant testing

207   In their seminal wavelet study, Torrence and Compo (1998) outline a method for

208   objectively identifying statistical significance in the wavelet power by comparing the wavelet

209   power spectra with a power spectra from a red noise process. Specifically, the observed time

210   series is fitted with an order 1 autoregressive (AR1, or red noise) model, and the WT is applied

211   to the AR1 time series. The power spectrum of the AR1 model provides the basis for the

212   statistical significance testing. Significance is determined if the power spectra are statistically

213   different using a chi-squared test.

214   Figure 2b shows significant (>= 95% confidence level) regions of wavelet power inside

215   black contours. Statistical significance indicates wavelet power that falls outside the time series

216   background statistical power based on an AR1 model of the time series. Statistical significance

217   of the wavelet power can be thought of as events in the wavelet domain. We define events as

218   regions of significant wavelet power outside the COI. Figure 2c displays the wavelet power for

219   the events in this time series. We emphasize that events defined in this way are a function of both

220   time and timescale and that, at a given time, events of different timescales can occur

221   simultaneously.

222   3.1.3. Step 1c. Sample observed events to an event-set relevant to analysis

223   Step 1b results in the identification of all events at all timescales and times. In this sub-

224   step, the event space is sampled to suit the particular evaluation. Torrence and Compo (1998)

225   offer two methods to smoothing the wavelet plot that can increase significance and confidence:

226   (i) averaging in time (over timescale) or (ii) averaging in timescale (over time). Because the goal

227   of this paper is to evaluate model timing errors over long simulation periods, we choose to

228   sample the event space based on averaging in timescale. Although for some locations there may

229   be physical reasons to expect certain timescales to be important (e.g., seasonal cycle of

230   snowmelt), the most important timescales at which hydrologic signals occur at a particular

231   location are not necessarily known *a priori*. Averaging events in timescale can provide a useful

232   diagnostic by identifying the dominant, or "characteristic", timescales for a given time series.

233   Averaging many events in timescale can filter noise and help reveal the expected timescales of

234   dominant variability corresponding to different processes or sets of processes.

235   In our analysis we seek to uncover the dominant event timescales and to evaluate modeled

236   timing errors on these. The following bullets articulate our methodological choices for

237   summarizing observed events:

238   • *Calculate the average event power in each timescale:* Considering only the statistically

239   significant areas of the observed wavelet spectrum, calculate the average power in each

240   timescale (Figure 2c, right panel). We point out that calculating the average power over

241   events is different than what is found by averaging across all time points, which doesn't

242   take statistical significance into consideration (Figure 2b, right panel).

243
244   • *Identify timescales of absolute and local maxima in time-average power:* After obtaining

245   the average event power as a function timescale (Figure 2c, right panel), the local and

246   absolute maximums for average event power can be determined. In the Onion Creek case,

247   there is a single maximum at 22 hours (grey dot in Figure 2c, right panel). The timescales

248   corresponding to the absolute and local maxima of the average power of the observed

249   time series are called the characteristic timescales used for evaluation. This is the first

250   subset of events: all events that fall within the characteristic timescales. For a single

251   characteristic timescale, contiguous events in time are called event clusters (horizontal

252   line in Figure 2d).

253
254   • *Identify events with maximum power in each event cluster:* For all timescales, we identify

255   the event with maximum power in each event cluster. This is the second event subset: all

256   events with maximum power in each cluster that falls within a characteristic timescale

257   (star in Figure 2d); these are called cluster maxima.

258
259 *3.2. Step 2. Calculate Timing Errors*

260   Step 1 identifies observed events by applying a wavelet transform to the observed time

261 series. To calculate the timing error of a modeled time series, we perform its cross wavelet

262 transform with the observed time series. Figure 3a shows the observed and modeled time series

263 used in our illustration of the methodology: the observed is the same isolated peak from Onion

264 Creek, TX, as in Figure 2a, and the synthetic modeled time series adds a prescribed timing error

265 of +5 hours to the observed. (Note that while the observed time series is identical in both, figures

266 2a and 3a have linear and log10 axes, respectively).

267 3.2.1. Step 2a. Apply cross-wavelet transform (XWT) to observations and simulations

268   The cross-wavelet transform (XWT) is performed between the observed and synthetic

269 time series. Given the WTs of an observed time series $W_n^X(s)$ and a modeled time series $W_n^Y(s)$,

270 the cross-wavelet spectrum can be defined as:

271 $$W_n^{XY}(s) = W_n^X(s)W_n^{Y*}(s),$$
272 where the asterix denotes the complex conjugate. The cross-wavelet power is defined as

273 $|W_n^{XY}(s)|$ and signifies the joint power of the two time series. The XWT between the Onion

274 Creek observations and the synthetic 5 hour offset time series is shown in Figure 3b, with power

275 represented by the color scale.

276   Similar to Step 1b of the WT, we can also calculate areas of significance for the XWT

277 power as shown by the black contour in Figure 3b. For the XWT, significance is calculated with

278 respect to the theoretical background wavelet spectra of each time series (Torrence and Compo,

279 1998). We define XWT events as points of significant XWT power outside the COI. XWT

280 events indicate significant joint variability between the observed and modeled time series.

281 Below, in step 2d, we employ XWT events as a basis for identifying hits and misses on observed

282 events for which the timing errors are calculated. Figure 3c shows the observed events (colors),

283 and the intersection between the observed and XWT events (dashed contour). As described later,

284 this intersection (inside dashed contour) is a region of hits where timing errors are considered

285 valid. Note that the early part of the observed events at shorter timescales is not in the XWT

286 events. This is because the timing offset in the modeled time series misses the early part of the

287 observed event for some timescales.

288 3.2.2. Step 2b. Calculate the cross-wavelet timing errors

289   For complex wavelets, such as the Morlet used in this paper, the individual WTs include

290 an imaginary component of the convolution. Together, the real and imaginary parts of the

291 convolution describe the phase of each time series with respect to the wavelet. The cross wavelet

292 transform combines the WTs in conjugate, allowing the calculation of a phase difference or

293 angle (radians) which can be computed as:

294
$$\phi_n^{XY}(s) = tan^{-1}\left[\frac{\Im(\langle s^{-1}W_n^{XY}(s)\rangle)}{\Re(\langle s^{-1}W_n^{XY}(s)\rangle)}\right]$$

295 Where $\Im$ is the imaginary and $\Re$ is the real component of $W_n^{XY}(s)$. The arrows in Figure 3b

296 indicate the phase difference for our example case, which are used to calculate the timing errors.

297 Note that these are calculated at all points in the wavelet domain.

298   The distance around the phase circle at each timescale is the Fourier period (hours). We

299 convert the phase angle into the timing errors (hours) as in Liu et al. (2011):

300
$$\Delta t_n^{XY}(s) = \phi_n^{XY}(s) * T/2\pi$$

301 where T is the equivalent Fourier period of the wavelet. Note that the maximum timing error

302 which can be represented at each timescale is half the Fourier period because the phase angle is

303 in the interval (-pi, pi). In other words, only timescales greater than 2E can accurately represent a

304 timing error E. Because the range of the *arctan* function is limited by ±pi, true phase angles

305 outside this range alias to angles inside this range. (For example, the phase angles 1.05 * pi and -

306 .95 * pi are both assigned to -.95*pi). Also note that when the wavelet transforms are

307 approximately antiphase, the computed phase differences and timing errors produce

308 corresponding bimodal distributions given noise in the data. Figure 3c shows phase aliasing in

309   the negative timing errors at timescales less than 10 hours, double the 5 hour synthetic timing

310   error we introduced. The bimodality of the phase and timing are also seen at the 10hr timescale

311   when the timing errors abruptly change sign (or phase by 2pi). We note the convention used is

312   that the XWT produces timing errors that are interpreted as "modeled minus observed", i.e.,

313   positive values mean the model occurs after the observed. Positive 5 hour timing errors in Figure

314   3c describe that the model is "late" compared to the observations as seen in the hydrographs in

315   the top panel (a).

316   3.2.3. Step 2c. Subset cross-wavelet timing errors to sampled observed events

317   Step 2b results in an estimate of timing errors for all times and timescales in the cross-

318   wavelet transform space. In our application, we are interested in the timing errors that correspond

319   to the identified sample of *observed* events, especially for the maximum power events in each

320   cluster for each characteristic timescale. In the synthetic Onion Creek example, the point of

321   interest in the wavelet transform of the observed timeseries, used to sample the timing errors

322   produced by the XWT, is shown by the grey star in Figure 3c.

323   The results for the synthetic Onion Creek example are summarized in Table 2. For the

324   identified characteristic timescale of 22 hours in the observed wavelet power (which had an

325   average WT power of 555,700 m^6/s^2 - Figure 2c right), there was 1 event cluster, and the

326   timing error at the cluster maximum was 5 hours and it occurred at hour 37 of the time series.

327   3.2.4. Step 2d. Filter Misses

328   The premise of computing a timing error between the observed and modeled time series

329   is that they share common events which can be meaningfully compared. In a two-way

330   contingency analysis of events, a "hit" refers to when the modeled time series reproduces an

331   observed event. When the modeled time series fails to reproduce an observed event, it is termed

332 a "miss". In the case of a miss, it does not make sense to include the timing error in the overall

333 assessment. Once the characteristic timescales of the observed event spectrum are identified and

334 event cluster maxima are located, timing errors are obtained at these locations in the XWT. In

335 this step, the significance of the XWT on these event cluster maxima is used to decide if the

336 model produced a hit or a miss for each point and to determine if the timing error is valid. As

337 previewed above, Figure 3c shows the observed events (colors) and the dashed contour shows

338 intersection between the observed and XWT events. Regions of intersection between observed

339 events and XWT events are considered model hits and observed events falling outside the XWT

340 events are considered misses. Because we constrain our analysis to observed events in the

341 wavelet power spectrum, we do not consider either of the remaining categories in a 2-way

342 analysis (false alarms and correct negatives). We note that a complete 2-way event analysis

343 could alternatively be constructed in the wavelet domain based on the Venn diagram of the

344 observed and modeled events without necessarily using the XWT. We choose to use the XWT

345 events because the XWT is the basis of the timing errors.

346 In the synthetic example of Onion Creek, a single characteristic timescale and event

347 cluster yields a single cluster maximum as shown by the star in Figure 3c. Because this star falls

348 both within the observed and XWT events, it is a hit and the timing error at that point is valid

349 (Table 2). For a longer time series, as seen in subsequent examples, a useful diagnostic and

350 compliment to timing error statistics at each characteristic timescale is the percent hits. When

351 summarizing timing error statistics for a timescale, we drop misses from the calculation and the

352 % hits indicates what portion of the time series was dropped (% misses = 100 - % hits). In our

353 tables we provided timing error statistics only for hits.

354 **4. Results**

355   In the previous section, we illustrate the method using an isolated peak and a prescribed

356   timing error. In this section, we demonstrate the method using NWM model simulations which

357   introduce greater complexity and longer time series. Finally, we show version-over-version

358   comparisons for 5-year simulations to illustrate the utility for evaluation.

359   4.1 Demonstration using NWM data

360   *Pemigewasset River, NH*

361   This example uses a three-month time series from the Pemigewasset River, NH, to examine

362   multiple peaks in the hydrograph (Figure 4a). By eye, it is fairly straightforward to pick out three

363   main peaks. From Step 1 of our method, the wavelet transform is applied to the observations

364   (Figure 4b, left and 4c, left), revealing up to three event clusters, depending on the characteristic

365   timescale examined (Figure 4d). When we plot the average event power by timescale (right of

366   Figure 4c), we see that there are nine relative maxima (small grey dots) – hence there are 9

367   characteristic scales for this example. The cluster maxima (grey stars) for each observed event

368   cluster are shown in Figure 4d.

369   Next, we compare the observed time series with the simulation from the NWM V1.2 (Figure

370   5a), and follow Step 2 of our method: a) apply the cross-wavelet transform (Figure 5b colors), b)

371   calculate the timing error for all observed events from the phase difference (Figure 5b arrows), c)

372   subset the timing errors to the observed cluster maxima (Figure 5c stars), and d) retain only

373   modeled hits (Figure 5c stars within the dashed contours). Table 3 is ordered by characteristic

374   timescales from highest to lowest average power; we only show the top 5 characteristic scales.

375   The absolute maximum of the time average event spectrum has a timescale = 24.8 hours; for

376   cluster one, the model is nearly 11 hours late and cluster two is early (-3.5 hours), both are hits,

377   and the average timing error is 3.5 hours late. However, for the next timescale (=27.8 hr), the

378 third cluster maximum is a miss, so its timing error is reported as a NA, and is not included in the

379 average. This miss can be seen in Figure 5c where the cluster 3 star falls just outside the XWT

380 events for the 27.8 timescale. Moreover, this miss can also be interpreted from the comparison of

381 the hydrographs in Figure 5a where the modeled third peak does not reasonably approximate the

382 magnitude of the observed peak. Interestingly, while it is a narrow miss at the shorter timescale

383 of 27.8 hours, the associated (3rd) cluster maxima at the next most powerful characteristic

384 timescale (33.1 hours) is a hit. This reflects that the hydrograph is insufficiently peaked for this

385 event but does have some of the observed, lower-frequency variability. Overall, the characteristic

386 timescale of 33.1 hours has timing results similar to the 27.8 hour timescale with the exception of

387 the third cluster maximum. This raises the question if these are distinct characteristic timescales.

388 In the Discussion and Conclusions section we discuss smoothing the time average event power

389 by timescale to address this issue.

390   The characteristic timescale with the 4th highest time-average power occurs at 111 hours,

391 which is a different order of magnitude, suggesting that this may have a different physical

392 process driving it. At this timescale, the model is late in both event clusters (10 and 16 hours).

393 Results are similar for the next timescale of 148 hours. We don't show results for the remaining

394 4 characteristic time scales with lower average power, since they have similar characteristic

395 timescale values and associated timing errors to what has already been shown.

396 *Taylor River, CO*

397

398   In this example, we examine a one-year time series from Taylor River, CO, that illustrates

399 hydrograph peaks driven by different processes. The Taylor River is in a mountainous area

400 where the spring hydrology is dominated by snowmelt runoff. Figure 6a shows the time series

401 from Taylor River, CO, where we can see the snowmelt runoff in spring and also several peaks

402   in summer, likely driven by summer rains. Figure 6b shows the WT, and illustrates how missing

403   data is handled: this results in additional COIs (muted colors) to account for the edge effects, and

404   areas of the COI are ignored in our analyses.

405   From the statistically significant events in the WT, we see the peak in the characteristic time

406   scales at 23.4 hours (right of Figure 6c), and there is another maxima at 99 and 118 hour

407   timescales. The process-based shift in dominant timescales is evident in the wavelet power

408   (Figure 6b and 6c): the 23.4 hour timescale is dominant before July 1, during snowmelt runoff,

409   and then shifts to the 99 and 118 hour timescales, relating to flows from summer rains. In Step 2,

410   we compare the observed time series with the simulation from the NWM V1.2 (Figure 7a); here

411   it is useful to zoom into the spring melt season time series (Figure 8), where we see that the

412   amplitude of the diurnal signal is too high, but it's hard to visually tell much about the timing

413   error. Next, the cross-wavelet transform (Figure 7b) and timing errors are calculated (Figure 7c).

414   The results are summarized in Table 4. Starting with the dominant 23.4 hour timescale, we see

415   that there are 11 clusters, that 73% (=8/11 cluster maxima) are hits, and that the model is

416   generally early (the mean is 6 hours early). For the 118 and 99 hour timescales, there are no hits.

417   This suggests that we are confident in the timing errors of the model for the diurnal snowmelt

418   cycle, and these timing errors can be used as guidance for model performance and model

419   improvements. However, the model does not successfully reproduce key variability during the

420   summer and timing errors are not valid at this timescale. This underscores the key point that

421   timing errors are timescale dependent, and can help diagnose which processes to target for

422   improvements.

423   4.2 Evaluating Model Performance
424

425   Finally, we show how the methodology can be used for evaluating performance changes

426   across NWM versions. We point out that none of the NWM version upgrades were targeting

427   timing errors, so these results just provide a demonstration. We use 5-year observed and modeled

428   time series at the three locations: Onion Creek, TX, Pemigewasset River, NH, and Taylor River,

429   CO.

430   For Onion Creek, Table 5 summarizes the results for the three most important timescales and

431   Figure 9 provides a graphical representation of these timing errors (hits only). For the dominant

432   29.5 hour timescale and for all model versions, there were 19 cluster maxima, 89.5% of which

433   were hits, with a median timing error of 1.4 hours early. However, the model shows

434   progressively earlier timing errors with increasing version (Figure 9). The results are similar for

435   the other two characteristic timescales.

436   For Pemigewasset River, Table 6 summarizes the results for the 3 most important timescales

437   and Figure 10 provides a graphical representation of the timing errors (hits only). At this

438   location, the median timing error improved with NWM V1.2, getting closer to zero. While the

439   distribution of the timing errors became less biased than the previous versions, it also became

440   wider (Figure 10). Over the timeseries, there were between 59 and 76 event clusters.

441   Interestingly, the hit rate for all timescales was best for NWM V1.1 though its timing errors are

442   broadly the worst. From NWM V1.0 to NWM V1.2, improvements to both hit rate and median

443   timing errors were obtained at all timescales.

444   For Taylor River, Table 7 summarizes the results for the 2 most important timescales. For the

445   characteristic timescale of 235 hours (~10 days) there are only 4 event clusters and each model

446   version has only 1 hit. The timing of this hit improves by roughly half its error from NWM V1.0

447   to NWM V1.2 in going from 16 to 9 hours. The 23.4 hour timescale has 41 event clusters with a

448 hit rate varying considerably by version. The median timing error is fairly consistent with

449 version, however, ranging from 6 to 7 hours early.

450 **6. Discussion and Conclusions**

451

452   In this paper, we develop a systematic, data-driven methodology to objectively identify

453 time series (hydrograph) events and estimate timing errors in large-sample, high-resolution

454 hydrologic models. The method was developed towards several intended uses: Primarily, it was

455 developed for model evaluation, so that model performance can be documented in terms of

456 defined standards. We illustrate this with the version-over-version NWM comparisons. Second,

457 it can be used for model development, whereby potential timing error sources can be diagnosed

458 (by timescale) and targeted for improvement. Related to this point, given the advantages of

459 calibrating using multiple-criteria (e.g., Gupta et al. 1998), timing errors could be used as part of

460 a larger calibration strategy. However, minimizing timing errors at one timescale may not

461 translate to improvements in timing errors (or other metrics) at other timescales. Wavelet

462 analysis has also been used directly as an objective function for calibration, although a difficulty

463 is in determining the similarity measure to use (e.g. Schaefli and Zehe 2009, Rathinasamy et al.

464 2014). Future research will investigate the application of the timing errors presented here for

465 calibration purposes. Finally, the approach can be used for model interpretation and forecast

466 guidance as estimating timing errors provides characterization of the timing uncertainty (i.e., for

467 a given timescale, the model is generally late or early) or confidence.

468   Given the fact that several subjective choices were made specific to our application and

469 goals, it is important to highlight that we have made the analysis framework openly available

470 (detailed in the code availability section below), so the method can be adapted, extended, or

471 refined by the community right away. We look at timing errors from an observed event-set

472   relevant to our analysis, but there are other ways to subset the events that might be more suitable

473   to other applications. For example, we focus on the event cluster maxima, but one could also

474   examine the event cluster means or the local maxima along time. Another alternative to finding

475   the event cluster maxima (i.e., for a given timescale) would be to identify the event with

476   maximum power in "islands of significance" across timescales, i.e., contiguous regions of

477   contiguous significance across both time and timescale. This approach would ignore that

478   multiple frequencies can be important at once. Also defining such islands is not straightforward.

479   A different approach could be desirable if one suspected non-stationarity in the characteristic

480   timescales over the time series. Then perhaps a moving average in timescale could be employed

481   to identify characteristic timescales. In our approach, we define the event set broadly. However,

482   it could be subset using streamflow thresholds (e.g. for flooding events) to compare events in the

483   wavelet domain with traditional peak-over-threshold events. For example, Figure 11 shows the

484   maximum streamflows for the event-set from the 5 year time series at Taylor River. This figure

485   shows that all events identified by the algorithm are not necessarily high flow events (i.e., the

486   maximum streamflow peaks are lower for the 23.4 hour timescale as compared to the 235.6 hour

487   timescale). To compare with traditional peak-over-threshold approaches, this event-set could be

488   filtered to include only events above a given threshold (i.e., events in both the wavelet and time

489   domains).

490   Another point that arises is how many characteristic timescales should be examined and

491   the similarity of adjacent characteristic timescales. In our method, we average the power in

492   timescales and identify characteristic scales to be at every absolute and relative maxima. As seen

493   in the illustrative examples, this can result in multiple characteristic scales, some of which can be

494   quite similar, suggesting that events at those scales are from similar or related processes. One

495 solution could be to smooth the average power by timescale, which would reduce the number of

496 local maxima, or to look at timing errors within a band of timescales. It is also important to note

497 that the characteristic scales are data-driven, so they will change with different lengths of

498 observed time series. Longer runs capture more events and should converge on the more

499 dominant timescales and events for a location. However, for performance evaluation,

500 overlapping time periods for observed and modeled time series are needed.

501   In our application of the WT, we follow Liu et al. (2011) and select the Morlet as the

502 mother wavelet. However, results are sensitive to the mother wavelet selected. Further discussion

503 of mother wavelet choices can be found in Torrence and Compo (1998) and in ElSaadani and

504 Krajewski (2017).

505   In summary, this paper provides a systematic, flexible, and computationally efficient

506 methodology for calculating model timing errors that is appropriate for model evaluation and

507 comparison, and is useful for model development and guidance. Based on the wavelet transform,

508 the method introduces timescale as a property of timing errors. The approach also identifies

509 streamflow events in the observed and modeled timeseries and only evaluates timing errors for

510 modeled events which are hits in a 2-way contingency analysis. Future work will apply the

511 approach to identify characteristic timescales across the United States, as well as to assess the

512 associated timing errors in the NWM.

513 **Code/Data Availability**

514 The code for reproducing the figures and tables in this paper are provided in the public github

515 repository https://github.com/NCAR/wavelet_timing with instructions for installing

516 dependencies. The core code used in the above repository is provided in the public "rwrfhydro"

517 R package https://github.com/NCAR/rwrfhydro. The code is written in the open-source R

518   language (R Core Team 2019) and builds off multiple, existing R packages. Most notably the

519   wavelet and cross-wavelet analyses are performed using the "biwavelet" package (Gouhier et al.

520   2018).

521   We emphasize that the analysis framework is meant to be flexible and adapted to similar

522   applications where different statistics may be desired. The figures created are specific to the

523   applications in this paper but provide a starting point for other work.

524   **Credit Author Statement**

525   ET and JLM collaborated to develop the methodology. ET led the results analysis, manuscript

526   preparation, and revisions. JLM led the initial idea for the work, the open source software, and

527   the visualizations.

528   **Competing interests**. The authors declare that they have no conflict of interest.
529

530   **Acknowledgements**:
531
532   The authors would like to thank Dave Gochis for useful discussions and Aubrey Dugger

533   for providing NWM data. We thank the NOAA/OWP and NCAR NWM team for its support of

534   this research. This research is funded by the NOAA Office of Water Prediction and the Joint

535   Technology Transfer Initiative grant 2018-0303-1556911. This material is based upon work

536   supported by the National Center for Atmospheric Research (NCAR), which is a major facility

537   sponsored by the National Science Foundation (NSF) under Cooperative Agreement No.

538   1852977.

539

540   **References**:
541
542   Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive
543   uncertainty estimation in a flood forecasting system. Water Resour. Res. 47, W07524,
544   doi:10.1029/2010WR009137.
545

546   Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values, Springer Ser. Stat.
547   Springer, London.
548
549   Daubechies, I., 1990. The wavelet transform time-frequency localization and signal analysis.
550   IEEE Trans. Inform. Theory 36, 961-1004.
551
552   ElSaadani, M., Krajewski, W. F., 2017. A time-based framework for evaluating hydrologic
553   routing methodologies using wavelet transform. *Journal of Water Resource and Protection*, *9*(7),
554   723–744.
555
556   Ehret, U., Zehe, E., 2011. Series distance-an intuitive metric to quantify hydrograph similarity in
557   terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System*
558   *Sciences*, *15*, 877–896. https://doi.org/10.5194/hess-15-877-2011
559
560   Gochis, D.J., M. Barlage, R. Cabell, M. Casali, A. Dugger, K. FitzGerald, M. McAllister, J.
561   McCreight, A. RafieeiNasab, L. Read, K. Sampson, D. Yates, Y. Zhang (2020). The WRF-
562   Hydro® modeling system technical description, (Version 5.1.1). *NCAR Technical Note.* 107
563   pages. Available online at:
564   https://ral.ucar.edu/sites/default/files/public/WRFHydroV511TechnicalDescription.pdf. Source
565   Code DOI:10.5281/zenodo.3625238
566
567   Gouhier, T.C., Grinsted, A., Simko, V., 2018. R package biwavelet: Conduct Univariate and
568   Bivariate Wavelet Analyses (Version 0.20.17), https://github.com/tgouhier/biwavelet.
569   Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G. F., 2009. Decomposition of the mean
570   squared error and NSE performance criteria: Implications for improving hydrological modelling.
571   Journal of hydrology, 377(1-2), 80-91.
572
573   Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V.,
574   2014. Large-sample hydrology: a need to balance depth with breadth. Hydrol. Earth Syst. Sci.
575   18, 463-477, https://doi.org/10.5194/hess-18-463-2014.
576
577   Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Towards improved calibration of hydrologic
578   models: multiple and non-commensurable measures of information. Water Resources Research
579   34(4): 751–763.
580
581   Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a
582   diagnostic approach to model evaluation. *Hydrological Processes*, *22*(March), 3802–3813.
583   https://doi.org/10.1002/hyp
584
585   Koskelo, A. I., Fisher, T. R., Utz, R. M., and Jordan, T. E., 2012. A new precipitation-based
586   method of baseflow separation and event identification for small watersheds (< 50 km2), J.
587   Hydrol., 450–451, 267–278, https://doi.org/10.1016/j.jhydrol.2012.04.055.
588
589   Lane, S.N., 2007. Assessment of rainfall–runoff models based upon wavelet analysis. Hydrol.
590   Processes 21, 586–607, https://doi.org/10.1002/hyp.6249.
591

592   Liu, Y., Liang, X.S., Weisberg, R.H., 2007. Rectification of the bias in the wavelet power
593   spectrum. J. Atmos. Oceanic Technol. 24, 2093–2102.
594

595   Liu, Y., Brown, J., Demargne, J., Seo, D. J., 2011. A wavelet-based approach to assessing timing
596   errors in hydrologic predictions. J. Hydrol. 397(3–4), 210–224.
597   http://doi.org/10.1016/j.jhydrol.2010.11.040
598

599   Luo, Y. Q., Randerson, J.T., Abramowitz, G., Bacour, C., Blyth, E., Carvalhais, N., … Orme, C.
600   E., 2012. A framework for benchmarking land models. Biogeosciences 9, 3857-3874,
601   https://doi.org/10.5194/bg-9-3857-2012.
602

603   McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., Rea, A., 2012. NHDPlus Version
604   2: user guide. National Operational Hydrologic Remote Sensing Center, Washington, DC.
605

606   Mei, Y. and Anagnostou, E. N., 2015. A hydrograph separation method based on information
607   from rainfall and runoff records, J. Hydrol., 523, 636–649,
608   https://doi.org/10.1016/j.jhydrol.2015.01.083.
609

610   Merz, R., Blöschl, G., and Parajka, J., 2006. Spatio-temporal variability of event runoff
611   coefficients, J. Hydrol., 331, 591–604, https://doi.org/10.1016/j.jhydrol.2006.06.008.
612

613   Newman, A.J., Mizukami, N., Clark, M.P., Wood, A.W., Nijssen, B., Nearing, G., 2017.
614   Benchmarking of a physically based hydrologic model. J. Hydrometeorol. 18, 2215–2225,
615   http://doi.org/10.1175/JHM-D-16-0284.1.
616

617   Niu, G.-Y., Yang, Z.-L., Mitchell, K.E., Chen, F., Ek, M.B., Barlage, M., Kumar, A., Manning,
618   K., Niyogi, D., Rosero, E., Tewari, M., Xia, Y., 2011. The community Noah land surface
619   model with multiparameterization options (Noah-MP): 1. Model description and evaluation with
620   local-scale measurements. J. Geophys. Res. 116, D12109, doi:10.1029/2010JD015139.
621

622   NOAA National Weather Service, 2012. NWS Manual 10-950. Definitions and General
623   Terminology. Hydrological Services Program, NWSPD 10-9,
624   http://www.nws.noaa.gov/directives/sym/pd01009050curr.pdf.
625

626   R Core Team, 2019. R: A language and environment for statistical computing. R Foundation for
627   Statistical Computing, Vienna, Austria, https://www.R-project.org/.
628

629   Rathinasamy, M., Khosa, R., Adamowski, J., Ch, S., Partheepan, G., Anand, J., & Narsimlu, B.,
630   2014. Wavelet-based multiscale performance analysis: An approach to assess and improve
631   hydrological models. *Water Resources Research*, *50*(12), 9721–9737.
632

633   Schaefli, B., Zehe, E., 2009. Hydrological model performance and parameter estimation in the
634   wavelet-domain. Hydrol. Earth Syst. Sci. 13, 1921-1936, https://doi.org/10.5194/hess-13-1921-
635   2009.
636

637   Seibert, S. P., Ehret, U., Zehe, E. 2016. Disentangling timing and amplitude errors in streamflow
638   simulations. Hydrology and Earth System Sciences, 20, 3745–3763. https://doi.org/10.5194/hess-
639   20-3745-2016
640
641   Torrence, C., Compo, G.P., 1998. A Practical Guide to Wavelet Analysis. Bull. Am. Meteorol.
642   Soc. 79(1), 61-78.
643
644   Veleda, D., Montagne, R., Araujo, M., 2012. Cross-wavelet bias corrected by normalizing
645   scales. J. Atmos. Ocean. Technol. 29, 1401-1408.
646
647   Weedon, G.P., Prudhomme, C., Crooks, S., Ellis, R.J., Folwell, S.S., Best, M.J., 2015.
648   Evaluating the performance of hydrological models via cross-spectral analysis: case study of the
649   Thames Basin, United Kingdom. J. Hydrometeorol. 16(1), 214–231. http://doi.org/10.1175/JHM-
650   D-14-0021.1.

Tables

Table 1. Nomenclature table of terms used in the manuscript

Term, Acronym	Synonyms	Units	Comments
time series	Input data	m^3/s for streamflow	We analyze streamflow observations and simulations, which are ordered by the time dimension (Figure 2a)
time		hours	Dimension of the input time series (x-axis on all Figure 2 panels).
timescale	period	hours	Dimension introduced at each time by the WT (y-axis on Figure 2b-d)
wavelet transform (WT)	wavelet power spectrum (result of the transform)	m^6/s^2	In this paper, we employ the continuous WT (Figure 2b) with scale normalized energy (Liu et al, 2007)
cone of influence (COI)	COI		Where wavelet analysis is affected by the wavelet extending beyond the time domain of the input (muted colors in Figure 2b)
event			We define events in terms of both time and timescales that are significant in the WT and outside the COI (Figure 2c)
characteristic timescale	dominant timescale	hours	We define characteristic timescales by local maxima in time-averaged, significant wavelet power (e.g. over events) (Figure 2d)
event cluster			For a single (e.g. characteristic) timescale, contiguous events in time (Figure 2d)
cross wavelet transform (XWT)	cross wavelet power spectrum (result of the transform)	Power: m^6/s^2 Phase: radians	The complex, cross-wavelet transform has properties of power and phase. Significance of the XWT can also be computed (e.g. Torrence and Compo, 1998) as shown in Figure 3b. XWT events are outlined in dashed line in figure 3c.
timing error		hours	Timing errors are calculated from the phase offset of the XWT (e.g. Liu, 2011) and have dimensions of both time and timescale. Several statistics of timing errors (over time) for characteristic timescales can be computed (Figure 2c).

Table 2. Summary of timing error results for cluster maxima for isolated peak and prescribed 5 hour offset from Onion Creek, TX.

Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	Cluster Maxima		
			Timing Error (hr)	Time (hr)	Hit?
22	555,700	1	5	37	TRUE

Table 3. For three-month time series from Pemigewasset River, NH, by characteristic timescale, summary of timing error results for cluster maxima that were hits using NWM v1.2.

Characteristic Timescale (hr)	Avg WT Power	Cluster	Timing Error (hr)	Hit?	Total Number of Clusters	% Hits	Avg Timing Error (hr)
24.8	82,800	1	10.7	TRUE	2	100	3.5
		2	-3.64	TRUE			
27.8	74,400	1	9.54	TRUE	3	67	2.8
		2	-3.99	TRUE			
		3	NA	FALSE			
33.1	73,100	1	8.41	TRUE	3	100	1.2
		2	-2.12	TRUE			
		3	-2.71	TRUE			
111	72,000	1	10.0	TRUE	2	100	13
		2	16.5	TRUE			
148	58,200	1	12.0	TRUE	2	100	14
		2	15.0	TRUE			

Table 4. For one year time series from Taylor River, CO, by characteristic timescale: summary of timing error results for the cluster maxima that were hits using NWM v1.2.

Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	% Hits	Timing Error (hr)		
				Min	Mean	Max
23.4	316	11	73	-8.2	-6.0	-3.6
118	93.1	2	0	NA	NA	NA
99.1	90.5	2	0	NA	NA	NA

Table 5. Summary of timing errors from cluster maxima that were hits for 5 year time series from Onion Creek, TX.

NWM Version	Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	% Hits	Median Timing Error (hr)
v1.0	29.5	2,843,000	19	89	-1.4
v1.1	29.5	2,843,000	19	89	-2.8
v1.2	29.5	2,843,000	19	89	-3.2
v1.0	17.5	2,672,000	26	92	-1.1
v1.1	17.5	2,672,000	26	88	-1.9
v1.2	17.5	2,672,000	26	92	-2.4
v1.0	58.9	1,578,000	14	79	-1.4
v1.1	58.9	1,578,000	14	79	-3.0
v1.2	58.9	1,578,000	14	79	-3.0

Table 6. Summary of timing errors from cluster maxima that were hits for 5 year time series from Pemigewasset River, NH.

NWM Version	Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	% Hits	Median Timing Error (hr)
v1.0	17.5	172,900	67	84	-2.7
v1.1	17.5	172,900	67	91	-2.8
v1.2	17.5	172,900	67	85	-0.2
v1.0	27.8	169,600	61	82	-3.9
v1.1	27.8	169,600	61	97	-4.2
v1.2	27.8	169,600	61	90	1.1
v1.0	31.2	169,500	59	86	-4.2
v1.1	31.2	169,500	59	95	-4.6
v1.2	31.2	169,500	59	93	1.6

Table 7. Summary of timing errors from cluster maxima that were hits for 5 year time series of Taylor River, CO.

NWM Version	Characteristic Timescale (hr)	Avg WT Power	Number of Clusters	% Hits	Median Timing Error (hr)
v1.0	236	263	4	25	-16
v1.1	236	263	4	25	-10
v1.2	236	263	4	25	-9.0
v1.0	23.4	250	41	68	-6.1
v1.1	23.4	250	41	44	-6.9
v1.2	23.4	250	41	56	-6.5

Figures
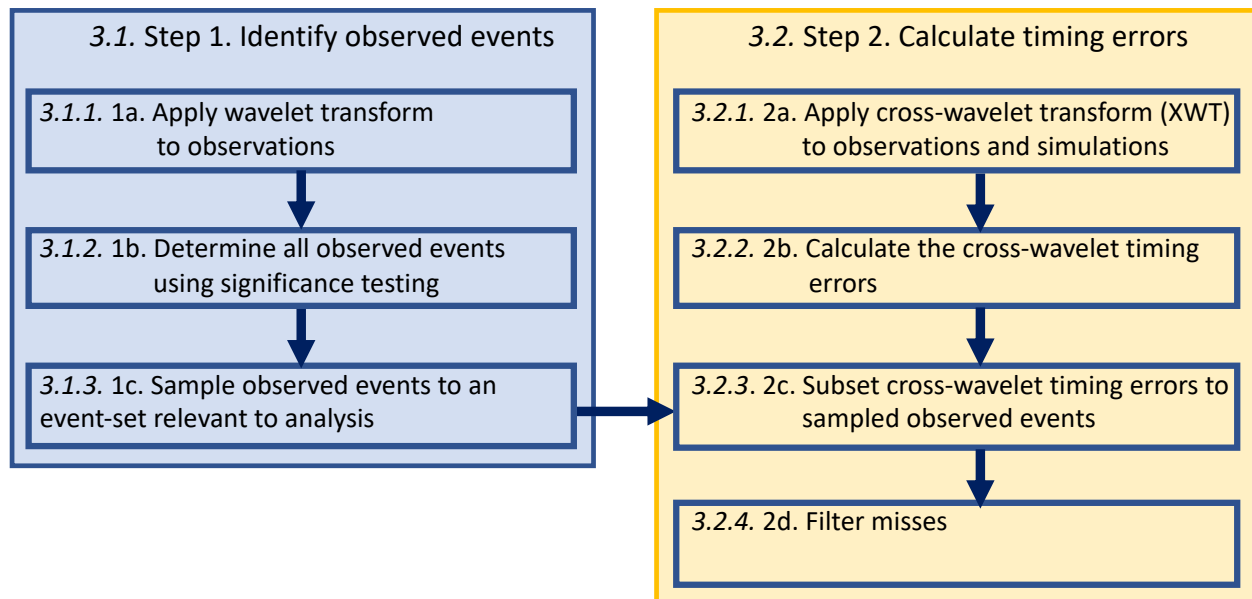
Figure 1. Flow chart of steps in the methodology; although Steps 1a-1b and Steps 2a-2b can happen in parallel, Step 2c needs to be preceded by Step 1c.
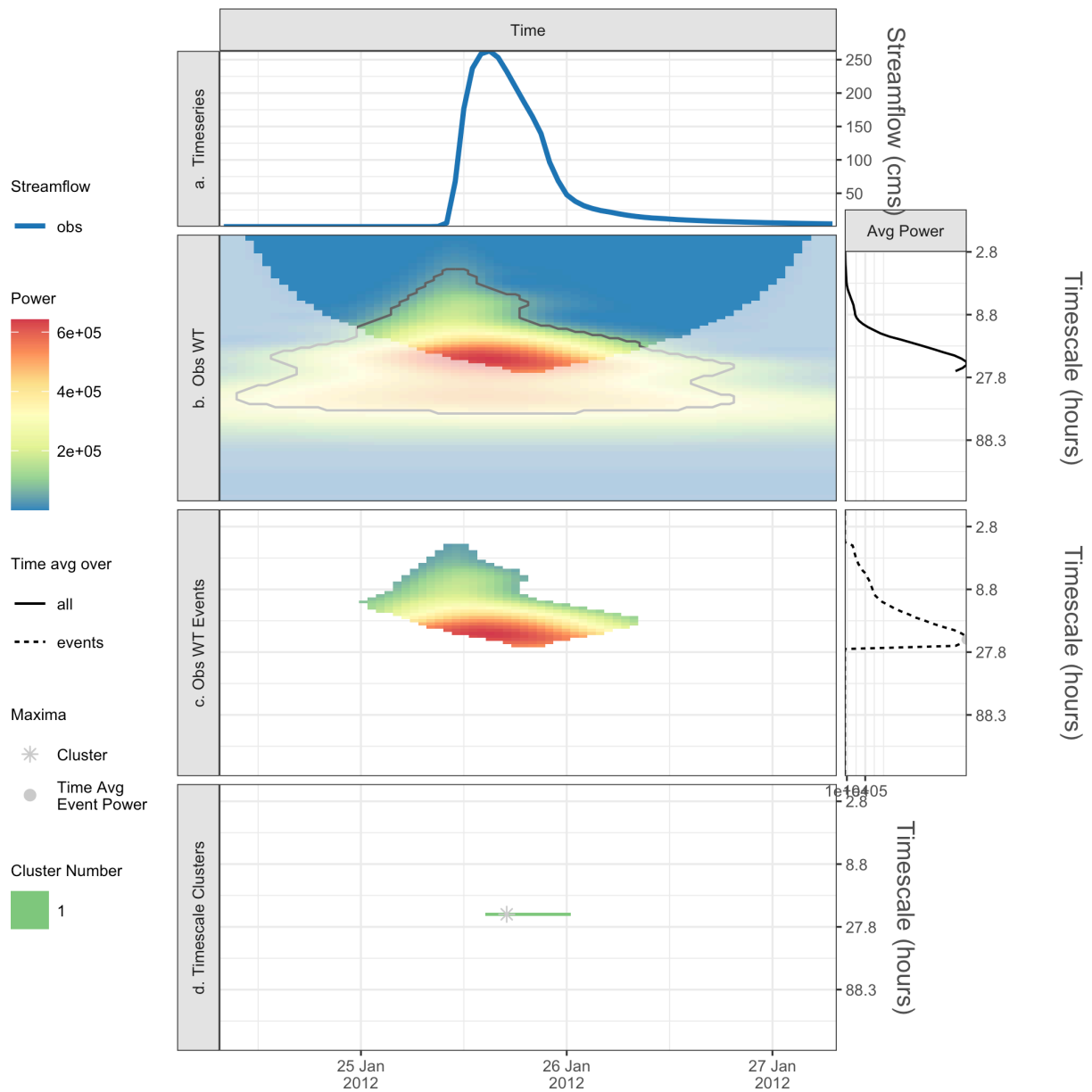
Figure 2. An isolated peak from Onion Creek, TX: (a) observed time series, (b) observed wavelet power spectrum (left) and average power by timescale for all points (right); (c) statistically significant wavelet power spectrum or events (left) and average power by time scale for all events with maxima shown by grey dots (right); (d) Characteristic scale event cluster (horizontal green line) and cluster maximum (star).
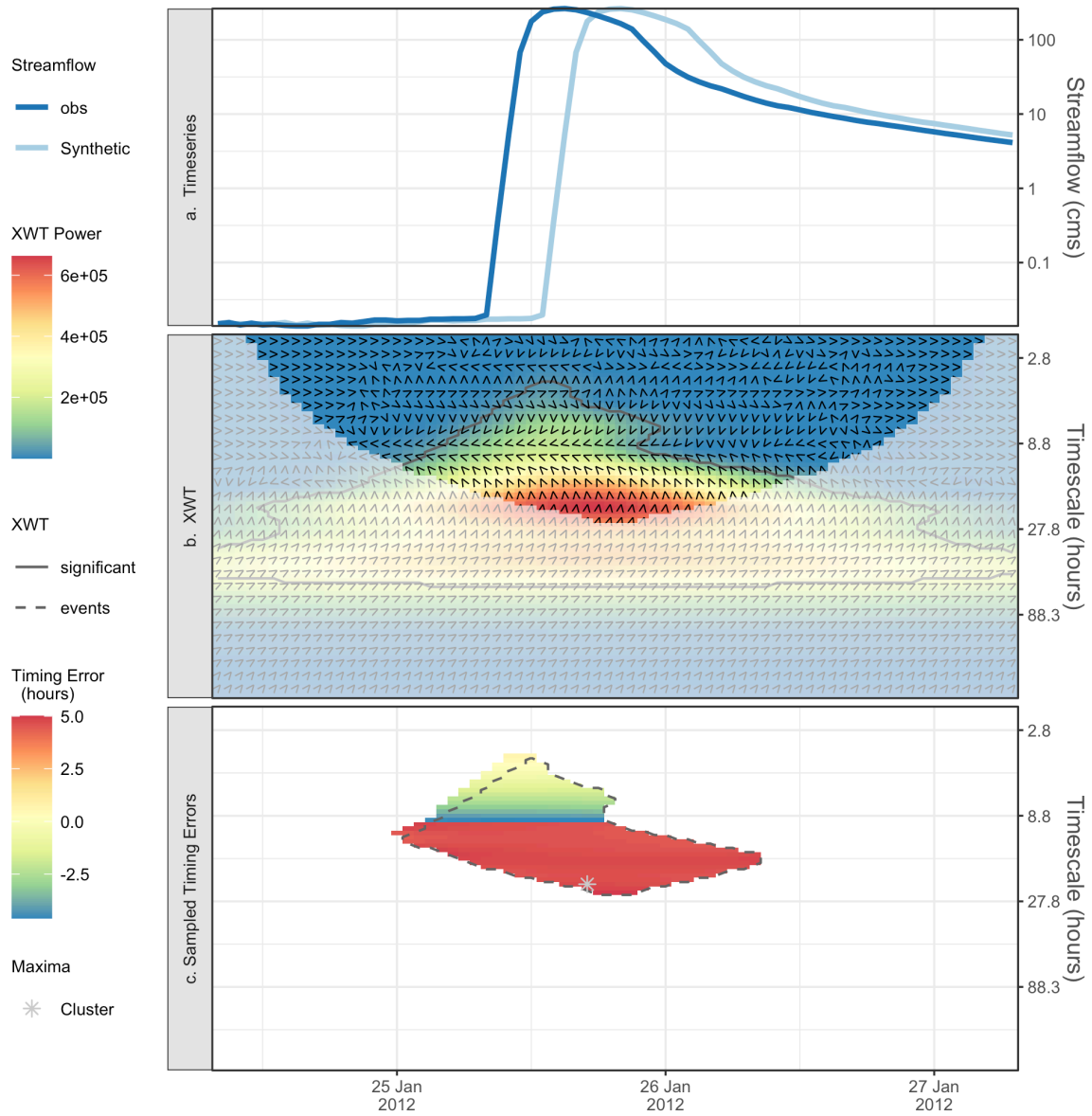
Figure 3. An isolated peak from Onion Creek, TX and a synthetic +5 hour offset: (a) observed and synthetic time series (note logged y-axis), (b) cross wavelet (XWT) power spectrum, phase angles (arrows), and XWT significance (grey line), (c) sampled timing errors for observed events (inside dashed contour indicates intersection of XWT events with observed events) and the grey star shows the cluster maximum from Figure 2d.
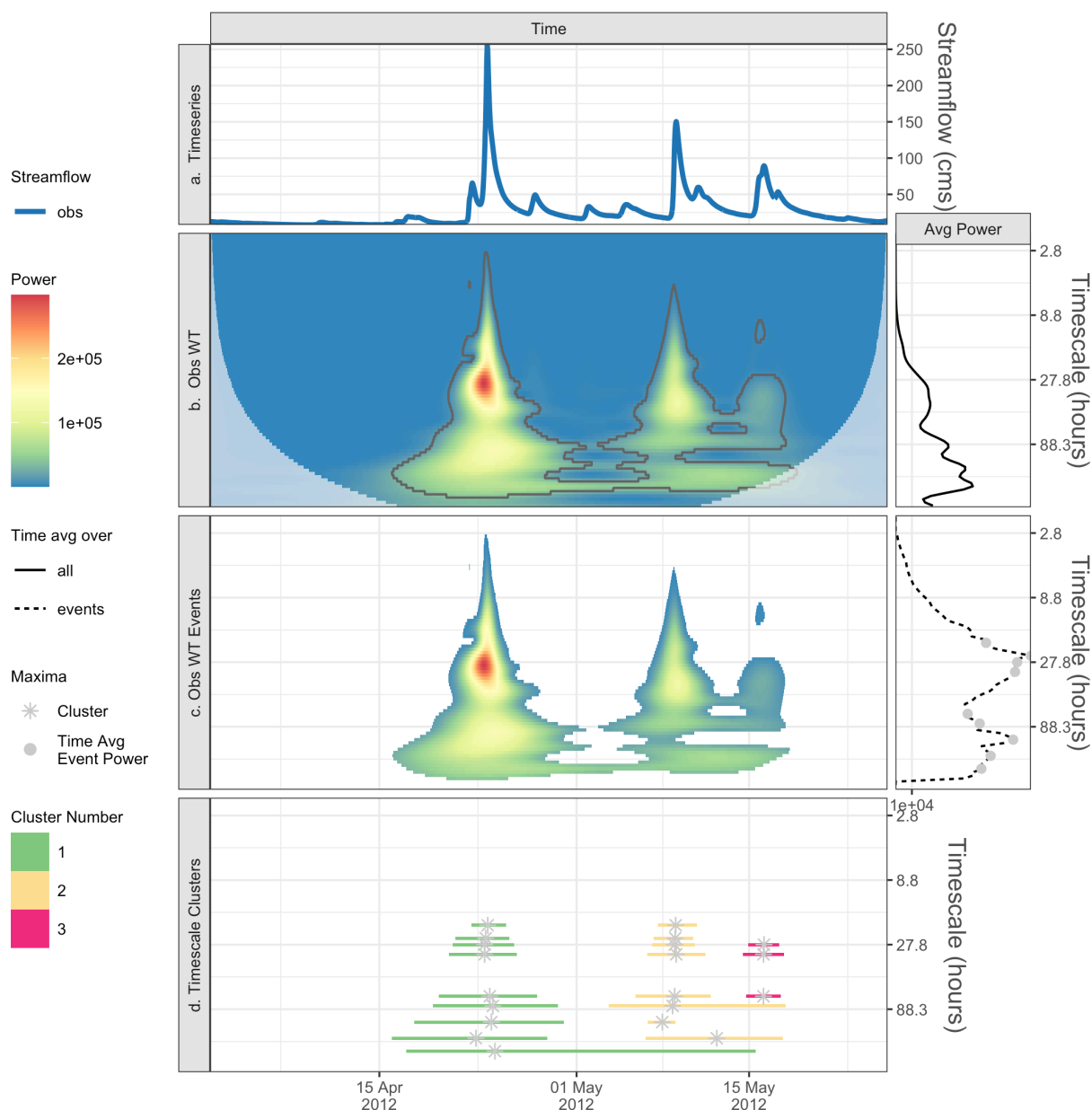
Figure 4. For three-month time series from Pemigewasset River, NH: (a) observed time series, (b) observed wavelet power spectrum (left) and average power by timescale for all points (right); (c) statistically significant wavelet power spectrum of events (left) and average power by time scale for all events with maxima shown by grey dots (right); (d) Characteristic scales event clusters (horizontal lines) and cluster maxima (grey stars).
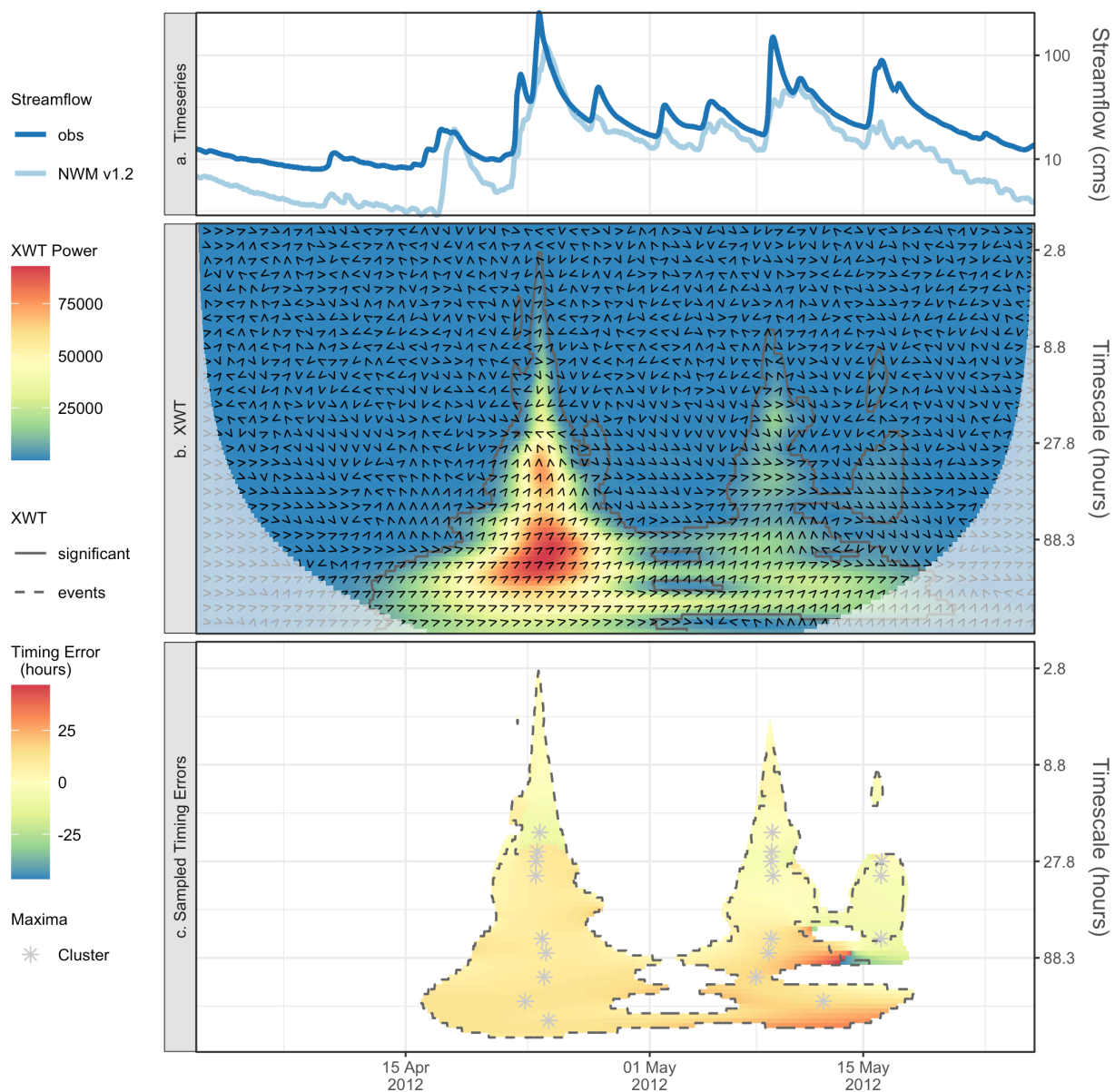
Figure 5. For three-month time series from Pemigewasset River, NH: (a) observed and simulated NWM time series (note logged y-axis), (b) cross wavelet (XWT) power spectrum (colors), phase angles (arrows), and statistically significant XWT events (solid contours), (c) sampled timing errors for observed events (inside dashed contour indicates intersection of XWT events with observed events) and cluster maxima (grey stars).
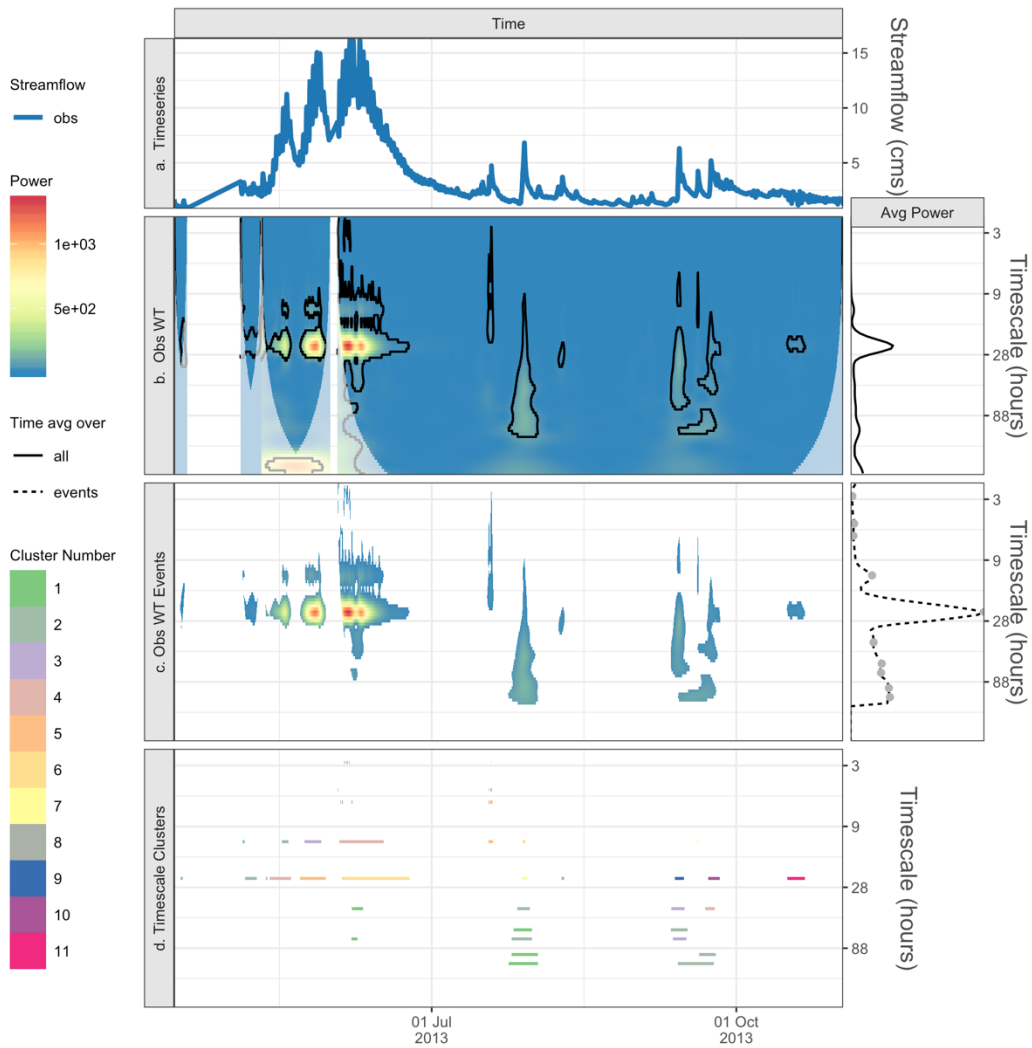
Figure 6 For one-year time series from Taylor River, CO: (a) observed time series, (b) observed wavelet power spectrum (left) and average power by timescale for all points (right); (c) statistically significant wavelet power spectrum or events (left) and average power by time scale for all events with maxima shown by grey dots (right); (d) Characteristic scales event clusters (horizontal lines).
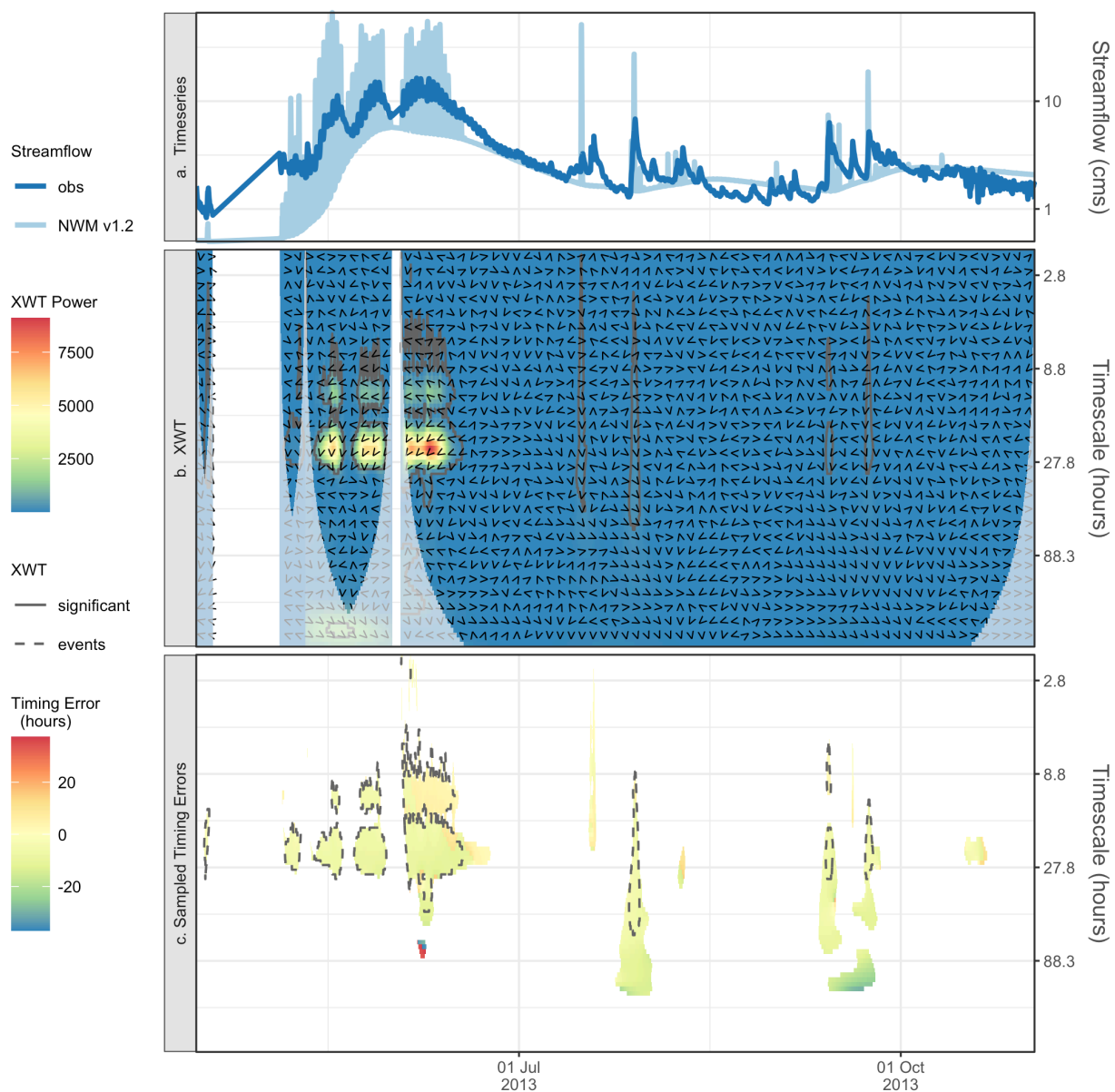
Figure 7. For one-year time series from Taylor River, CO: (a) observed and simulated NWM time series (note logged y-axis), (b) cross wavelet (XWT) power spectrum (colors), phase angles (arrows), and statistically significant XWT events (solid contours), (c) sampled timing errors for observed events (inside dashed contour indicates intersection of XWT events with observed events).
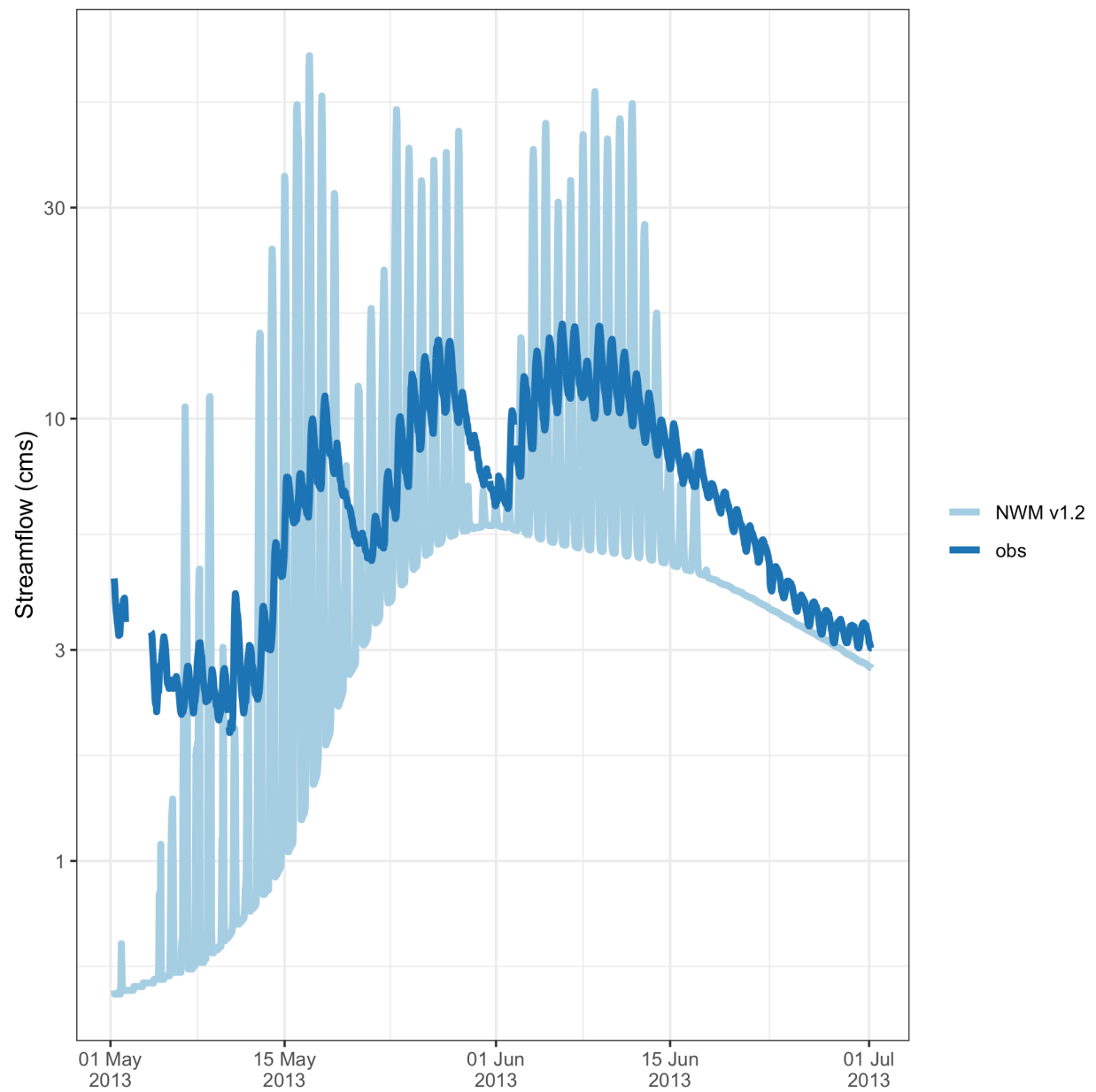
Figure 8. Zoom-in to spring runoff of one-year time series for Taylor Park, CO: observed and simulated NWM time series.
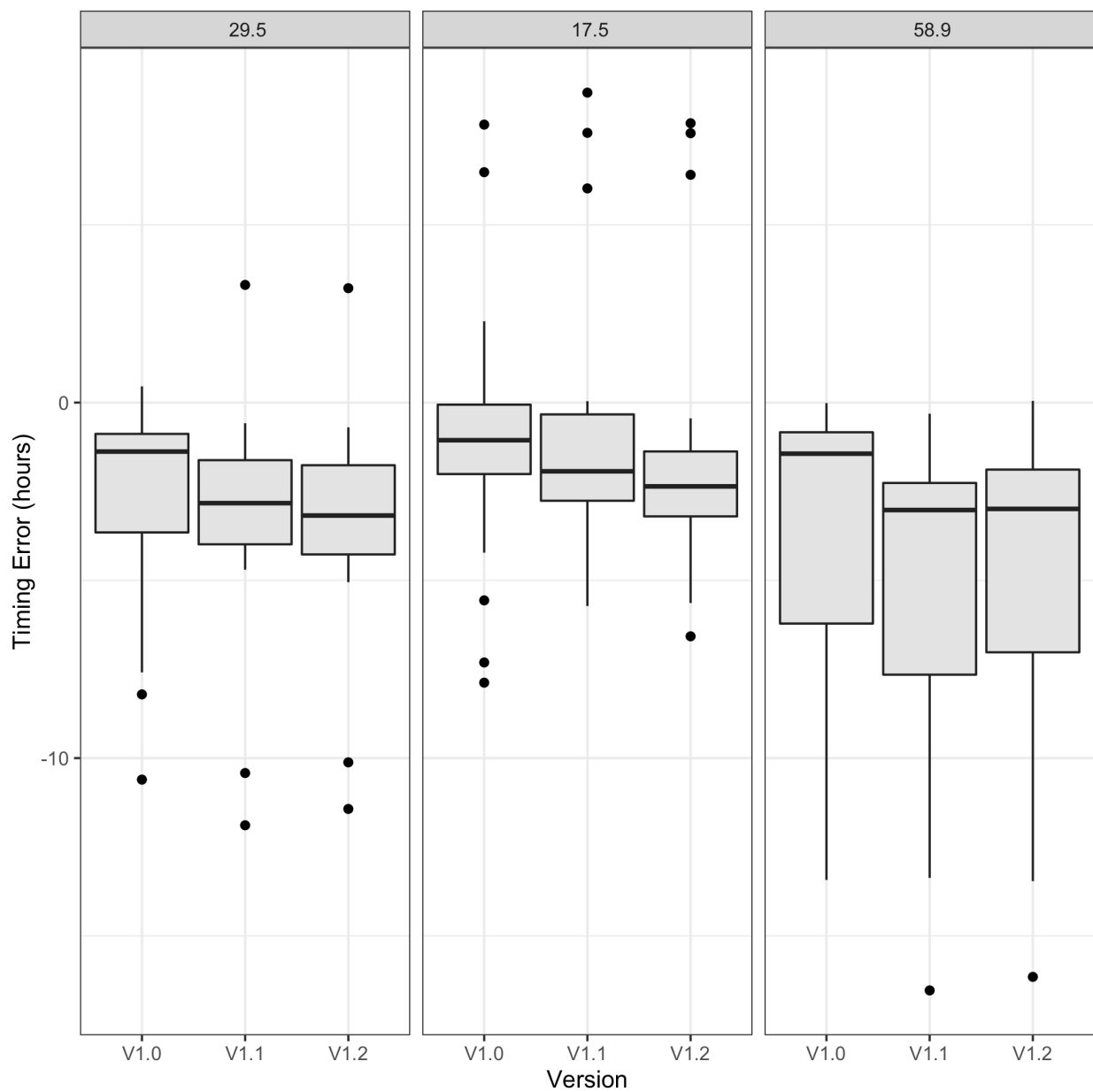
Figure 9. Five year time series from Onion Creek, TX: Comparing cluster maxima timing error distributions for top three characteristic timescales (see panel title) across NWM versions.
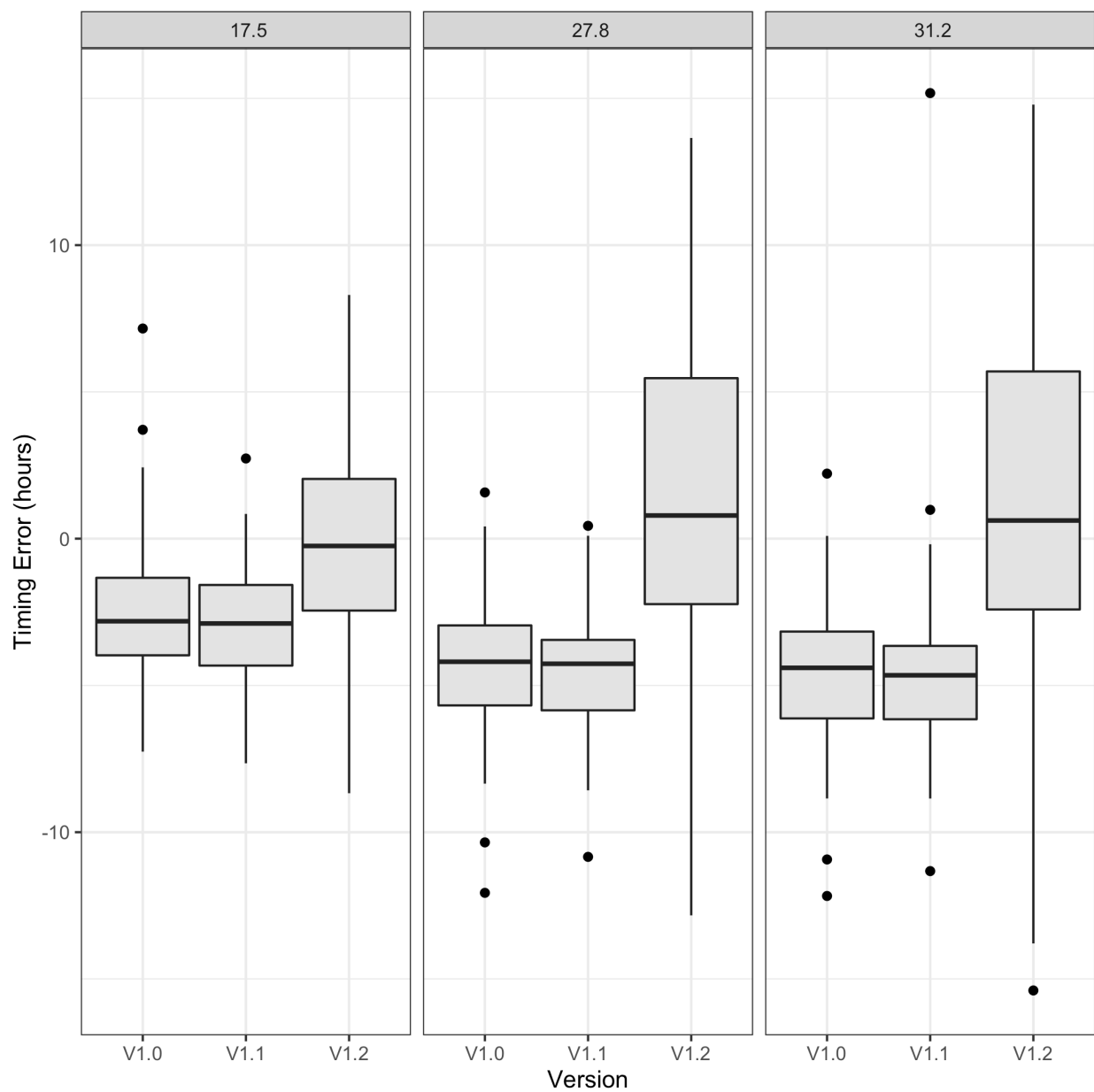
Figure 10. Five year time series from Pemigewasset River, NH: Comparing cluster maxima timing error distributions for top three characteristic timescales (see panel title) across NWM versions.
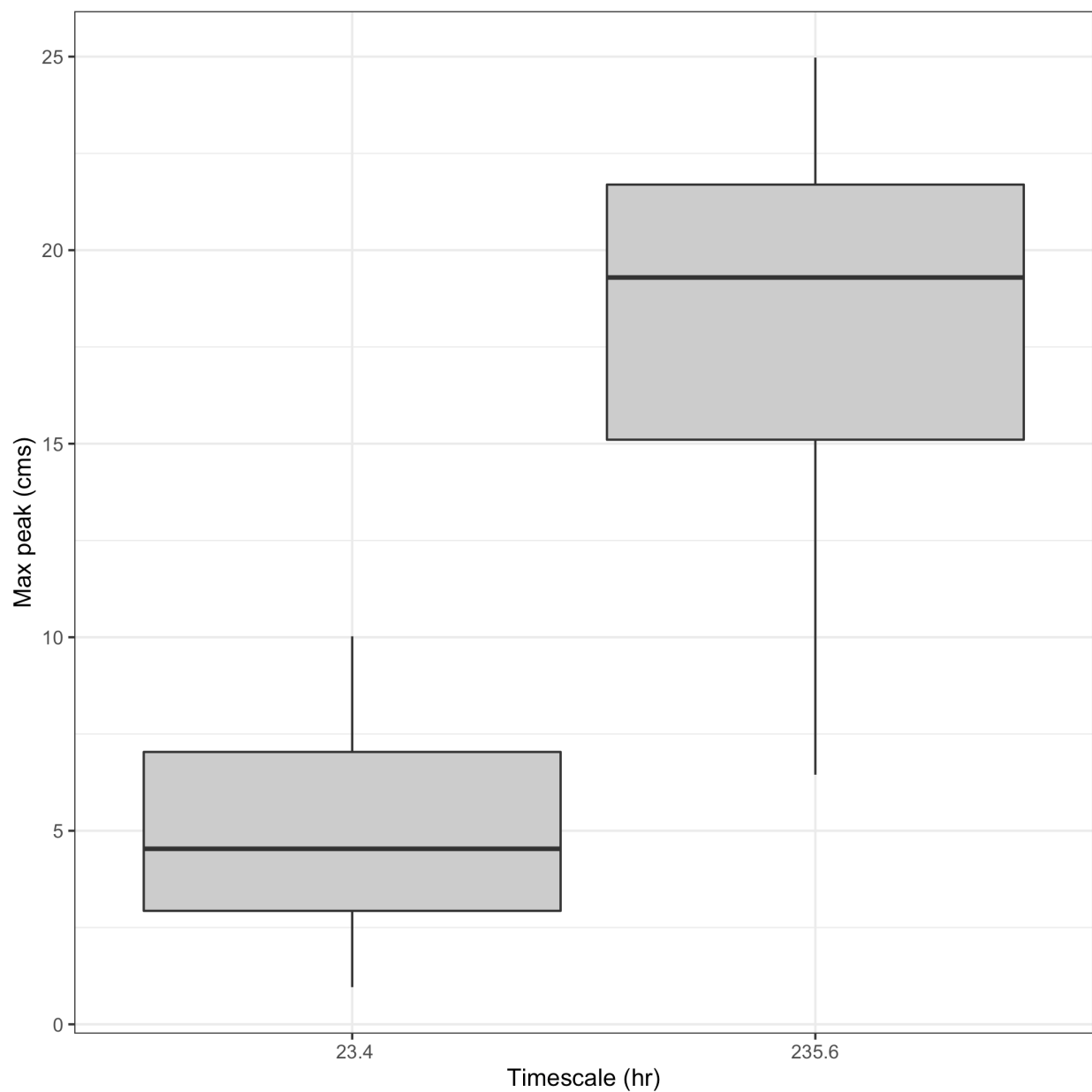
Figure 11. Five year time series from Taylor River, CO: for top two characteristic timescales, maximum streamflow peak distributions for each event (using cluster maxima) in cubic meters per second (cms).