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

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

Abstract

Streamflow timing errors (in the units of time) are rarely explicitly evaluated, but are useful for model evaluation and development. Wavelet-based approaches have been shown to reliably quantify timing errors in streamflow simulations, but have not been applied in a systematic way that is suitable for model evaluation. This paper provides a step-by-step methodology that objectively identifies events, and then estimates timing errors for those events, in a way that can be applied to large-sample, high-resolution predictions. Step 1 applies the wavelet transform to the observations, and uses statistical significance to identify observed events. Step 2 utilizes the cross-wavelet transform to calculate the timing errors for the events identified in Step 1; this ~~. This step e approach also~~ includes the ~~a quantification of the confidence in~~diagnostic of model event “hits”, and timing errors are only assessed for hits~~if the model “missed” observed evend and if the timing error estimatesshould not be considered~~. The methodology is illustrated using real and simulated stream discharge data from several locations to highlight key method features. The method groups event timing errors by dominant timescales, which can be used to identify the potential processes contributing to the timing errors and the associated model development needs.~~-~~ For instance, timing errors that are associated with the diurnal melt cycle are identified. The method is also useful for documenting and evaluating model performance in terms of defined standards. This is illustrated by showing version-over-version performance of the National Water Model (NWM) in terms of timing errors.

1. Introduction

Common verification metrics used to evaluate streamflow simulations are typically aggregated measures of model performance, e.g., the Nash Sutcliffe Efficiency (NSE) and the related root mean square error (RMSE). Although typically used to assess errors in amplitude, these statistical metrics include contributions from errors in both amplitude and timing (Ehret and Zehe 2011), making them difficult to use for diagnostic model evaluation (Gupta et al. 2008). Furthermore, common verification metrics are calculated using the entire time series, whereas timing errors require comparing localized features or events in the data. This paper focuses explicitly on event timing error estimation, which is not routinely evaluated, despite its potential benefit for model diagnostics (Gupta et al. 2008) and practical forecast guidance (Liu et al. 2011).

The fundamental challenge with evaluating timing errors is identifying what constitutes as an "event" in the two time series being compared. Identifying events is typically subjective, time consuming, and not practical for large-sample hydrological applications (Gupta et al. 2014). A variety of baseflow separation methods, ranging from physically-based to empirical, have been developed to identify hydrologic events (see Mei and Anagnostou 2015 for a summary), though many of these approaches require some manual inspection of the hydrographs. Merz et al. (2006) put forth an automated approach, but it requires a calibrated hydrologic model, which is a limitation in data poor regions. Koskelo et al. (2012) developed a simple, empirical approach that only requires rainfall and runoff time series, but is limited to small watersheds and daily data. Mei and Anagnostou (2015) introduce an automated physically-based approach, which is demonstrated for hourly data, though one caveat is that basin events need to have a clearly detectable recession period. ~~Most~~ Additional methods ~~for identifying events~~ have focused on

~~flooding events. One common approach to~~ identifying flooding events ~~is to~~ usinge peak-over-threshold methods. The thresholds used for such analyses are often either based on historical percentiles (e.g., the 95th percentile) or on local impact levels (river stage), such as the National Weather Service (NWS) flood categories (NOAA National Weather Service, 2012). Timing error metrics are often calculated from the peaks of these identified events. For example, the Peak Time Error, or its derivative the Mean Absolute Peak Time Error, requires matching observed and simulated event peaks, and calculating their offset (Ehret and Zehe 2011). While this may be straightforward visually, it can be difficult to automate; some of the reasons for this are discussed below.

Difficulties arise using thresholds for event identification. For example, exceedances can cluster if a hydrograph vacillates above and below a threshold, begging the question: Is it one or multiple events? Which peak should be used for the assessment? In the statistics of extremes, declustering approaches can be applied to extract independent peaks (e.g., Coles 2001), but this reductionist approach may miss relevant features. For instance, if background flows are elevated for a longer period of time before and after the occurrence of these "events", the threshold-based analysis identifies features of the flow separately from the primary hydrologic process responsible for the event. If one focuses just on peak timing differences in this example, that timing error may only apply to some small fraction of the total flow of the larger event which happens mainly below the threshold. Further, for overall model diagnosis that focuses on model performance for all events, not just flood events, variable thresholds would be needed to account for different kinds of events (e.g., a daily melt event versus a convective precipitation event).

Using a threshold-approach to identify events and timing error assessment, Ehret and Zehe (2011) develop an intuitive assessment of hydrograph similarity, the Series Distance. This algorithm is later improved upon by Siebert et al. (2016). The procedure matches observed and simulated segments (rise or recession) of an event, and then calculates the amplitude and timing errors, as well as the frequency of event agreement. The Series Distance requires smoothing the time series, identifying an event threshold, and selecting a time range to consider two segments matching.

Liu et al (2011) developed a wavelet-based method for estimating model timing errors. Although wavelets have been applied in many hydrologic applications such as model analysis (e.g. Lane 2007; Weedon et al. 2015; Schaefli and Zehe 2009, Rathinasamy et al. 2014) and post-processing (Bogner and Kalas 2007; Bogner and Pappenberger 2011), Liu et al. were the first to use it for timing error estimation. Liu et al. (2011) apply a cross-wavelet transform technique to streamflow time series for 11 headwater basins in Texas. Timing errors are estimated for medium- to -high- flow "events" that are determined a priori by threshold exceedance. They use synthetic as well as real streamflow simulations to test the utility of the approach. They show that the technique can reliably estimate timing errors, though they conclude that it is less reliable for multi-peak or consecutive "events" (defined qualitatively). ElSaadani and Krajewski (2017) followed the cross-wavelet approach used by Liu et al (2011) to provide similar analysis and further investigate the effect of the choice of mother wavelet on the timing error analysis. Ultimately, they recommended that in the situation of multiple, adjoining flow peaks the improved time localization of the Paul wavelet might justify its poorer frequency localization compared the Morlet wavelet.

Liu et al. (2011) provide a starting point for the work in this paper where we develop two new bases for their method: 1) objective event identification for timing error evaluation and 2) the use of observed events as the basis for the model timing error calculations. The latter is important for “model benchmarking”, i.e., the practice of evaluating models in terms of defined standards (e.g., Luo, et al. 2012; Newman et al. 2017). Here, the use of observed events provides a baseline by which to evaluate changes and to compare multiple versions or experimental designs.

This paper provides a methodology for using wavelet analysis to quantify timing errors in hydrologic simulations. Our contribution is a systematic approach that integrates 1) statistical significance to identify events with 2) a basis for timing error calculations independent of model simulations (i.e., benchmarking). We apply our method to timing error evaluation of high-resolution streamflow prediction. The paper is organized as follows: Section 2 describes the observational and simulated data used. ~~provides an overview of the conceptual approach of using wavelets to identify events and estimate timing errors, and~~ Section 3 provides the detailed methodology of using wavelets to identify events and estimate timing errors in a synthetic example. ~~In Section 4, we describe the software and data, as well as provide a simple illustration of the method using real and simulated streamflow data.~~ In Section 4~~5~~, we ~~provide results~~demonstrate the method using real and simulated streamflow data for several use cases, and then illustrate the application of the method for version-over-version comparisons. ~~, including select examples to highlight features of the method and version-over-version comparisons.~~ Section ~~6~~5 is the discussion and conclusions, including how specific methodological choices may vary by application.

2. Data

The application of the methodology is illustrated using real and simulated stream discharge (streamflow, m3/s) data ~~from~~at three U.S. Geological Survey (USGS) stream gage locations ~~representing~~in ~~three~~ different geographic regions: Onion Creek at US Highway 183, Austin, Texas, for the South Central region (Onion Creek, TX; USGS site number 08159000), Taylor River at Taylor Park, Colorado, for the Intermountain West (Taylor River, CO; USGS site number 09107000), and Pemigewasset River at Woodstock, New Hampshire, for New England (Pemigewasset River, NH; USGS site number 01075000). We use the USGS instantaneous observations averaged on an hourly basis.

NOAA's National Water Model (NWM, https://www.nco.ncep.noaa.gov/pmb/products/nwm/) is an operational model that produces hydrologic analyses and forecasts over the continental United States (CONUS) and Hawaii (as of version 2.0). The model is forced by downscaled atmospheric states and fluxes from NOAA's operational weather models. Next, the NoahMP (Niu et al 2011) land surface model calculates energy and water states and fluxes. Water fluxes propagate down the model chain through overland and subsurface (soil and aquifer representations) water routing schemes to reach a stream channel model. The NWM applies the three parameter Muskingum-Cunge river routing scheme to a modified version of the NHD-Plus version 2 (McKay et al. 2012) river network representation (Gochis et al 2020).

In this study, NWM simulations are taken from each version's retrospective runs (https://docs.opendata.aws/nwm-archive/readme.html). These are continuous simulations (not cycles) run for the period October 2010 to November 2016 and forced by the National Land Data Assimilation System (NLDAS)-2 product as atmospheric conditions. The nudging data assimilation was not applied in these runs ~~either~~. We use NWM discharge simulations from versions V1.0, V1.1, and V1.2 (not all version may be publicly available).

To apply the methodology, we note that the observed and simulated datasets must be paired (overlapping). The methodology developed in this paper is implemented in the R language and is made publicly available, as detailed in the code availability section at the end of the manuscript.

2. Conceptual Overview

This section provides the technical description of the methodology, and the steps can be seen in an accompanying flowchart (Supplemental Figure 1).Before going into technical details of the Method (Section 3), we provide a conceptual overview of the approach of using wavelets to identify events and estimate timing errors. We provide a nomenclature table (Supplemental Table 1) of key terms relevant to the approach. The wavelet transform (WT) expands the dimensionality of the original time series by introducing the timescale (or period) dimension and returns power as a function of both time and timescale (e.g. Torrence and Compo, 1998). This is illustrated in Figure 1: the streamflow time series (panel a) is expanded into a 2-dimensional wavelet power spectrum (panel b). Where traditional model errors, such as the aforementioned RMSE or NSE, reduce the information of the time series to a single statistic, wavelet analysis expands the input signal and provides information on the dominant timescales of the time series at each time. Wavelet analysis can therefore detect localized signals in time series (Daubechies 1990), including hydrologic time series, which are often irregular or aperiodic (i.e., events may be isolated and don't regularly repeat) or non-stationary. We note that in many wavelet applications, timescale is referred to as "period". To emphasize that our study is more focused on irregular events and less on periodic behavior of time series, we use the term "timescale". The wavelet transform is the foundation of the view in this paper that events have characteristics of both time and timescale. Timing errors, calculated from events defined this way, therefore have dimensions of both time and timescale as well.

In their seminal wavelet study, Torrence and Compo (1998) outline a method for objectively identifying statistical significance in the wavelet transform. We adopt this approach and define "events" in the observed time series via statistical significance of the wavelet power spectrum. The details are provided in the next section, however Figure 1 illustrates that the events in the input time series (panel a) are defined as regions of the wavelet power spectrum shown in panel b: events are inside the black contours (>= 95% confidence level) but not inside the cone of influence (regions where the colors are muted, this is explained in detail in Section 3). The wavelet power spectrum is only shown for the events in panel c. Events defined in this way are a function of both time and timescale. Note that at a given time, events of different timescales can occur simultaneously. What one may subjectively interpret as a single event in the input time series is generally quantified by this definition as multiple coincident events at a variety of timescales each with a different power (e.g. Figure 1, panel c). Although for some locations there may be physical reasons to expect certain timescales to be important (e.g., seasonal cycle of snowmelt), the most important scales at which hydrologic signals occur at a particular location are not necessarily known a priori. The wavelet power can be examined across events to identify the most dominant, or what we call "characteristic" timescales for a given time series; the procedure for this is detailed later in the technical methodological section (Section 3.1.3). This approach to event detection is objective, data-driven, and portable across diverse locations, which is important for large-sample hydrologic applications. We point out that in the objective identification of events, we are not limited to flooding events. Rather, events are defined more broadly: an event is when the wavelet power falls outside its standard statistical power. This can be further subset into flooding events if desired.

Once observed events are identified by the method, we can calculate timing errors between observed and simulated time series. The cross-wavelet timing error approach of Liu et al (2011) is used, but we restrict our calculation of timing errors to the aforementioned regions of statistically significant wavelet power in the observations; i.e., we calculate timing errors in terms of *observed* events (Figure 1c). Because both the phase (timing error) and the significance of the cross wavelet transform (XWT) computed between the observed and modeled time series depends on the modeled time series, we use the observed event definition (Figure 1c) in the calculation of the timing errors to provide a common, consistent basis independent of the models evaluated (i.e., benchmarking). The portions of the observed wavelet spectrum used for comparison may further be restricted depending on the analysis goals.

3. Methodology for evaluating event timing errors

This section provides the technical description of the methodology, and the steps can be seen in an accompanying flowchart (Supplemental Figure 1). This section provides the description of the methodology using wavelets to identify events and estimate timing errors. The steps can be seen in an accompanying flowchart (Figure 1) and nomenclature table (Table 1), which defines key terms of the approach. To facilitate understanding, tthe steps are illustrated accompanied by an application of the methodology to an observed time series of an isolated peak in Onion Creek, TX (Figure 2a), (figure 2a) and the synthetic modeled time series which is identical to the observation time seriesuniformly but shifted 5 hours in to the future (figure 3a, note the log sealescale).

3.1. Step 1. Identify observed events

The first step towards evaluating timing errors is to identify a set of observed events for which the timing error should be calculated. We break this step into three sub-steps: 1a. Apply the wavelet transform to observations, 1b. Determine all observed events using significance testing, and 1c. Sample observed events to an event-set relevant to analysis. To facilitate understanding, the steps are accompanied by an application of the methodology to an observed time series of an isolated peak in Onion Creek, TX (Figure 2a).

3.1.1. Step 1a. Apply wavelet transform to observations

First, we apply the continuous wavelet transform (WT) to the observed time series. The main steps and equations for the WT are provided here, though the reader is referred to Torrence and Compo (1998) and Liu et al. (2011) for more details.

Before applying the WT, a mother wavelet needs to be selected. In Torrence and Compo (1998), they discuss the key factors that should be considered when choosing the mother wavelet. There are four main considerations, including (i) orthogonal or nonorthogonal, (ii) complex or real, (iii) width, and (iv) shape. In this study, we follow Liu et al. (2011) in selecting the nonorthogonal and complex Morlet wavelet:

$$\psi(n) = \pi^{-1/4} e^{i w_0 n} e^{-n^2/2},$$

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

Once the mother wavelet is selected, the WT is applied to a time series x_n, where n goes from n=0 to n=N-1, with a time step of δt. The WT is the convolution of the time series with the mother wavelet that has been scaled and normalized:

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

where n' is the localized time in [0, N-1], s is the scale parameter, and the asterix indicates the complex conjugate of the wavelet function. The wavelet power is defined as $|W_n^2|$ which represents the squared amplitude of an imaginary number when a complex wavelet is used as in this study. We use the bias corrected wavelet power ~~spectrum~~ (Liu et al. 2007; Veleda et al. 2012), which ensures ~~spectral peaks are~~ power is comparable across timescales. We also identify a maximum timescale *a priori* that corresponds to our application. We select 256 hours (~10 days), but this number could be higher or lower for other applications and there are no real penalties for using too high a maximum (lower than the annual cycle).

The wavelet transform (WT) expands the dimensionality of the original time series by introducing the timescale (or period) dimension. Wavelet power ~~and returns power as~~ is also a function of both time and timescale (e.g. Torrence and Compo, 1998). This is illustrated in Figure ~~1~~2: the streamflow time series (panel a) is expanded into a 2-dimensional (2-D) wavelet power spectrum (panel b). ~~Where traditional model errors, such as the aforementioned RMSE or NSE, reduce the information of the time series to a single statistic, wavelet analysis expands the input signal and provides information on the dominant timescales of the time series at each time.~~ Wavelet analysis can ~~therefore~~ detect localized signals in time series (Daubechies 1990), including hydrologic time series, which are often irregular or aperiodic (i.e., events may be isolated and don't regularly repeat) or non-stationary. ~~We note that in many wavelet applications, timescale is referred to as "period". To emphasize that our study is more focused on irregular events and less on periodic behavior of time series, we use the term "timescale". The wavelet transform is the foundation of the view in this paper that events have characteristics of both time and timescale. Timing errors, calculated from events defined this way, therefore have dimensions of both time and timescale as well.~~We note that in many wavelet applications, timescale is referred to as "period" and this axis is indeed the Fourier period in our plots. However, to emphasize that our study is more focused on irregular events and less on periodic behavior of time series, we use the term “timescale” to denote Fourier period (and not wavelet scale). We provide an overview of the main steps and equations for the wavelet transform here, though the reader is referred to Torrence and Compo (1998) and Liu et al. (2011) for more details.

Before applying the WT, a mother wavelet needs to be selected. In Torrence and Compo (1998), they discuss the key factors that should be considered when choosing the mother wavelet. There are four main considerations, including (i) orthogonal or nonorthogonal, (ii) complex or real, (iii) width, and (iv) shape. In this study, we follow Liu et al. (2011) in selecting the nonorthogonal and complex Morlet wavelet:

$$\psi(n) = \pi^{-1/4} e^{i w_0 n} e^{-n^2/2},$$

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

Once the mother wavelet is selected, the WT is applied to a time series x_n, where n goes from n=0 to n=N-1, with a time step of δt. The WT is the convolution of the time series with the mother wavelet that has been scaled and normalized:

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

where s is the scale parameter, the asterix indicates the complex conjugate of the wavelet function. The wavelet power is defined as $|W_n^2|$. We use the bias corrected wavelet power spectrum (Liu et al. 2007; Veleda et al. 2012), which ensures spectral peaks are comparable across timescales. We also identify a maximum timescale that corresponds to our application. We select 256 hours (~10 days), but this number could be higher or lower for other applications and there are no real penalties for using too high a maximum (lower than the annual cycle). Because we are applying the WT to a finite time series, there are timescale-dependentBecause we are applying the WT to a finite time series, there are timescale-dependent errors at the beginning and end times of the power spectrum, where the entirety of the wavelet at each scale is not fully contained within the time series. This region of the WT is referred to as the cone of influence or COI (Torrence and Compo, 1998). ~~errors at the beginning and end times of the power spectrum. These are referred to as the cone of influence or COI (Torrence and Compo, 1998).~~ Figure 2b illustrates the COI as the regions where the colors are muted; ~~W~~we ignore all results within the COI in this study. ~~The details are provided in the next section, however Figure 1 illustrates that the events in the input time series (panel a) are defined as regions of the wavelet power spectrum shown in panel b: events are inside the black contours (>= 95% confidence level) but not inside the cone of influence (regions where the colors are muted, this is explained in detail in Section 3).~~

We make several additional notes on the wavelet power and its representation in the figures. The units of the wavelet power are those of the time series variance (m6/s2 for streamflow) and it is natural to want to cast the power in a physical light or relate it to the time series variance. Indeed, the power is often normalized by the time series variance when presented graphically. However, it must be noted that the wavelet convolved with the time -series frames the resulting power in terms of itself at a given scale. Wavelet power is a (normalized) measure of how well the wavelet and the time series match at a given time and scale. The power can only be compared to other values of power resulting from a similarly constructed WT. There are various transforms that can be applied to aid graphical interpretation of the power (log, variance scaling), but the utility of these often depends on the nature of the individual time series analyzed. For simplicity, we plot the raw bias-rectified wavelet power in this paper.

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

Once the WT is applied, the 2-dimensional (2-D) wavelet power spectra shows how the features of the time series vary with both time and timescale. In their seminal wavelet study, Torrence and Compo (1998) outline a method for objectively identifying statistical significance in the wavelet transform.power We adopt this approach and define “events” in the observed time series via statistical significance of the wavelet power spectrum. To identify areas of significance, we apply Torrence and Compo’s (1998) approach thatby comparesing the waveletWT power spectra with a power spectra from a red noise process. Specifically, the observed time series is fitted with an order 1 autoregressive (AR1, or red noise) model, and the WT is applied to the AR1 time series. The power spectruma of the AR1 model provides the basis for the statistical significance testing. Significance is determined if the power spectra are statistically different using a chi-squared test with 95% confidence.

We apply this here, and Figure 2b shows significant (>= 95% confidence level) regions of wavelet powerillustrates the events, which are inside the black contours (>= 95% confidence level) but not inside the COI. SStatistical significance indicates an “event” at a given time and timescale: that is, the wavelet that power that falls outside the time seriesits standard background statistical power based on an AR1 model of the time series. Statistical significance of the wavelet power can be thought of as events in the wavelet domain. We define events as regions of significant wavelet power outside the COI. Figure 2c displays the wavelet power for the events in this this time seriesThe result is the set of all events, i.e., each event is a combination of time and timescale (i.e., locations on the 2-D grid). The wavelet power spectrum is only shown for the events in Figure 2panel c. We emphasize that eEvents defined in this way are a function of both time and timescale . Noteand that, at a given time, events of different timescales can occur simultaneously. What one may subjectively interpret as a single event in the input time series is generally quantified by this definition as multiple coincident events at a variety of timescales each with a different power (e.g. Figure 1, panel c).

We refer to contiguous regions of statistical significance (in time and timescale) as “event clusters” (note that no statistical clustering is performed).

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

Step 1b results in the identification of all events at all timescales and times. In this sub-step, the event space is sampled to suit the particular evaluation. Torrence and Compo (1998) offer two methods to smoothing the wavelet plot that can increase significance and confidence: (i) averaging in time (over timescale) or (ii) averaging in timescale (over time). Because the goal of this paper is to evaluate model timing errors over long simulation periods, we choose to sample the event space based on dominant timescales in the time-averaged observed wavelet spectraaveraging in timescale. Although for some locations there may be physical reasons to expect certain timescales to be important (e.g., seasonal cycle of snowmelt), the most important timescales at which hydrologic signals occur at a particular location are not necessarily known *a priori*. Averaging events in timescale can provide a useful diagnostic be used toby identifyingy the most dominant, or what we call “characteristic”, timescales for a given time series, which is useful for model diagnostics. Averaging many events in timescale can fillter noise and help reveal the expected timescales of dominant variability corresponding to different processes or sets of processes. If one suspected nonstationarity in the timescales dominant variability over the timeseries, a different approach such as a moving average in timescale could be employed.The assumption is that identifiable sets of processes of interest are distinct in timescale, and that averaging over many events will reveal its expected value.

In our analysis we seek to uncover the dominant event timescales and to evaluate modeled timing errors on these. The following bullets articulate our methodological choices for summarizing observed events~~what was followed here~~~~For our application we choose to further sub-sample the observed wavelet spectra by selecting, for each characteristic timescale, the most powerful event within each event cluster. This is articulated in the following bullets~~:

- *Calculate the average event power ~~across~~ in each timescale:* Considering only the statistically significant areas of the observed wavelet spectrum, calculate the average power ~~across~~ in each timescale (Figure 2c, right panel~~over time~~). We point out that calculating the average power over events is different than what is found by averaging across all time points, which doesn't take statistical significance into consideration (Figure 2b, right panel). ~~.~~

- *Identify timescales of absolute and local maxima in time-~~-~~average power ~~maxima~~:* ~~By plotting the~~ After obtaining the average event ~~average~~ power ~~versus the~~ as a function timescale (Figure 2c, right panel), the local and absolute maximums for average event power can be determined. ~~(grey dots in Figure 2c, right); i~~In the Onion Creek case, there is a single maximum at 22 hours (grey dot in Figure 2c, right panel).~~.~~ The timescales corresponding to the absolute and local maxima of the average power of the observed time series are called the characteristic timescales ~~of the observed wavelet spectrum~~used for evaluation. This is the first subset of events: all events that fall within the characteristic time~~-~~scales. ~~-~~For a single characteristic timescale, contiguous events in time are called event clusters (horizontal line in Figure 2d).

- *Identify events with maximum power ~~for~~ in each event cluster:* For all timescales, ~~As previously mentioned, events can also be grouped into "event clusters", that is, contiguous significant areas~~ ~~For each event cluster,~~ ~~. We can use this to further sample from the event set created in the last bullet: across each characteristic timescale,~~ we identify the event with maximum power in ~~for~~ each event cluster. This is the second event subset: all events with maximum power ~~for~~ in each cluster that falls within a characteristic timescale~~.~~ (star in Figure 2d); these are called cluster ~~maximums~~maxima ~~(maxs)~~.

3.2. Step 2. Calculate Timing Errors

Step 1 identifies ~~characteristics (cluster maxima) of~~ observed events by applying a wavelet transform to the observed time series. To calculate the timing error of a modeled time series, we perform its cross wavelet transform with the observed time series~~, as detailed in this section~~. Figure 3a shows the observed and modeled time ~~time~~ series used in our illustration of the methodology: ~~To illustrate Step 2, we use~~ the observed is the same isolated peak from Onion Creek, TX ~~(Figure 2a)~~, as in Figure 2a, and the synthetic modeled time -series adds ~~and add~~ a prescribed timing error of +5 hours to the observed~~to every point in the original time series (Figure 3a) to create a synthetic time series~~. (Note that while the observed time series is identical in both, figures 2a and 3a have linear and log10 axes, respectively).

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

~~For Step 2, we use the same Onion Creek, TX, peak from Figure 1a, and add a prescribed timing error of +5 hours to every point in the original time series (Figure 2a) to create a synthetic time series.~~ The cross-wavelet transform (XWT) is performed between the observed and synthetic time series. ~~We perform the cross-wavelet transform between the observed and~~

synthetic time series (Figure 2b). Given the WTs of an observed time series $W_n^X(s)$ and a modeled time series $W_n^Y(s)$, the cross-wavelet spectrum can be defined as:

$$W_n^{XY}(s) = W_n^X(s) W_n^{Y*}(s),$$

where the asterix implies denotes the complex conjugate. The cross-wavelet power is defined as $|W_n^{XY}(s)|$. and signifies the joint power of the two time series. The XWT between the Onion Creek observations and the synthetic 5 hour offset time series isare shown in Figure 3b, with power represented by the color scale.

Similar to Step 1b of the WT, we can also calculate areas of significance for the XWT power as shown by the black contour in Figure 3b. For the XWT, significance is calculated with respect to the theoretical background wavelet spectra of each time series (Torrence and Compo, 1998). We define XWT events as points of significant XWT power outside the COI. XWT events indicate significant joint variability between the observed and modeled time series. Below, in step 2d, we employ XWT events as a basis for identifying hits and misses on observed events for which the timing errors are calculated. Figure 3c shows the intersection of the observed events (colors) and the XWT events (dashed contour). As described later, this intersection (inside dashed contour) is a region of hits where timing errors are considered valid. Note that the early part of the observed events at shorter timescales is not in the XWT events. This is because the timing offset in the modeled time series misses the early part of the observed event in a way that depends on timescale.

Similar to Step 1b of the WT, we can also calculate the areas of significance for the XWTblack . In the next section, widentifyinghits and misses for the events for which the timing errors are calculatedthe confidence in theare quantified by looking at the percent hits, i.e., These are not the same as the areas of significance for the WT. The significant areas of the XWT vary with each simulation, and are therefore not useful for evaluation on their own. Nevertheless, we

~~are interested in the overlap between the significant areas of the observed WT and the significant areas of the cross-wavelet transform, and this is used to quantify our confidence in the timing error estimate. We discuss this further in Step 2d.~~dashed

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

For complex wavelets, such as the Morlet used in this paper, the individual WTs include an imaginary component of the convolution. Together, the real and imaginary parts of the convolution describe the phase of each time series with respect to the wavelet. The cross wavelet transform combines the WTs in conjugate, allowing the calculation of a phase difference or angle (radians) which can be computed as

~~To calculate the timing errors, we first compute the phase angle of the cross-wavelet spectrum. The phase angle gives the phase difference and can be computed as~~:

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

Where $\Im$ is the imaginary and $\Re$ is the real component of $W_n^{XY}(s)$. The arrows in Figure 3b indicate the phase difference for our example case, which are used to calculate the timing errors. Note that these are calculated at all points in the wavelet domain.

The distance around the phase circle at each timescale is the Fourier period (hours). We convert the phase angle into the timing errors (hours) ~~We convert the phase angle into the timing error~~ as in Liu et al. (2011):

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

where T is the equivalent Fourier period of the wavelet.

. Note that the maximum timing error which can be represented at each timescale is half the Fourier period because the phase angle is in the interval (-pi, pi). In other words, only timescales greater than 2E can accurately represent a timing error~~,~~ E. Because the range of the *arctan* function is limited by ±pi, true phase angles outside this range alias to angles inside this range. (For example, the phase angles 1.05 * pi and -.95 * pi are both assigned to -.95*pi~~t~~). Also note that when the wavelet transforms are approximately antiphase, the computed phase differences and timing errors produce corresponding bimodal distributions given noise in the data. Figure 3c shows phase aliasing in the negative timing errors at timescales less than 10 hours, double the 5 hour synthetic timing error we introduced. The bimodality of the phase and timing are also seen at ~~that~~ the 10hr timescale when the timing errors abruptly change sign (or phase by 2pi). We note the convention used is that the XWT produces timing errors that are interpreted as “modeled minus observed”, i.e., ~~so that~~ positive values mean the model occurs after the observed. Positive 5 hour timing errors in Figure 3c describe that the model is “late” compared to the observations as seen in the hydrographs in the top panel (a).

~~We convert the phase angle into the timing error as in Liu et al. (2011):~~

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

~~where T is the equivalent Fourier period of the wavelet.~~

~~The arrows in Figure 3b indicate the phase offset, which are used to calculate the timing errors.~~

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

Step 2b results in an estimate of timing errors for all times and timescales in the cross-wavelet transform space. In our application, we are interested in the timing errors that correspond to the identified sample of *observed* events, especially for ~~events at the characteristic timescales (the first event-set in step 1c) and for~~ the maximum power events in each cluster ~~(the second~~

event-set in step 1c). At for each characteristic timescale, this provides a timing errors at each event cluster's maximum value. In the synthetic Onion Creek example, the points of interest in the wavelet transform of the observed timeseries, used to sample the timing errors produced by the XWT, areis shown by the grey star in Figure 3c. The latter provides a single timing error for each event cluster max at each characteristic timescale, which could be used in a post-processing step to provide a cluster-by-cluster timing correction, if desired.

In Figure 3c, the timing error estimates show that for timescales greater than 10 hours, we get back the prescribed timing error of 5 hours, i.e., the scale must be at least double the timing error. The results for the synthetic Onion Creek example are summarized in Table 2.; Ffor the identified characteristic timescale of 22 hours in the observed wavelet power, (which had an average WT power of 555,700676598,000 m^6/s^2 - (from Figure 2c right), there was 1 event cluster, and the timing error forat the cluster maximum was 5 hours (and it, which occurred at hour 37 of the time series). maximum

3.2.4. Step 2d. QuantifyFilter Misses Percent Hits

The premise of computing a timing error between the observed and modeled time series is that they share common events which can be meaningfully compared. In a two-way contingency analysis of events, a "hit" refers to when the modeled time series reproduces an observed event. When the modeled time series fails to reproduce an observed event, it is termed a "miss". In the case of a miss, it does not make sense to include the timing error in the overall assessment. Because the timing errors are calculated from the XWT, we choose to diagnose hits and misses based on the significance of the XWT. Once the characteristic timescales of the observed event spectrum are identified and event cluster maxima are located, timing errors are obtained at these locations in the XWT. In this step, the significance of the XWT on these event cluster maxima is used to decide if the model produced a hit or a miss for each point and to determine if the timing error is valid. ~~For a single cluster max, such as shown in Figure 3c, the XWT significance is either True or False, the point is either a hit or a miss.~~ As previewed above, Figure 3c shows ~~shows~~ the observed events (colors) and the XWT events (dashed contour). Regions of intersection between observed events and XWT events are considered model hits~~the intersection of the observed events (colors) and the XWT events (dashed contour)~~ and ~~.~~observed events falling outside the XWT events are considered misses. Because we constrain our analysis to observed events in the wavelet power spectrum, we do not consider either of the remaining categories in a 2-way analysis (false alarms and correct negatives). We note that a complete 2-way event analysis could alternatively be constructed in the wavelet domain based on the Venn diagram of the observed and modeled events without necessarily using the XWT. W~~As mentioned, w~~e choose to use the XWT events because the XWT is the basis of the timing errors.

In the synthetic example of Onion Creek, ~~This is a region of hits.~~ ~~For~~ a single characteristic timescale and event cluster yields a single cluster max~~, such as~~ as shown by the star in Figure 3c. Because this star falls both within the observed and XWT events, it is a hit and the timing error at that point is valid (Table 2)~~, the XWT significance is either True or False, the cluster max (star) point is either a hit or a miss.~~ ~~Table 2 summarizes the results of the timing error analysis for this synthetic example. We can see the prescribed 5 hour offset is recovered by the calculation and that the timing error is valid because the observed event was reproduced by the model (a hit).~~ For a longer time series, as seen in subsequent examples, a useful diagnostic and compliment to timing error statistics at each characteristic timescale is the percent hits. When summarizing timing errors statistics for a timescale, we drop misses from the calculation and the % hits indicates what portion of the time series was dropped (% misses = 100 - % hits).

In our tables we provided timing error statistics only for hitsthis way as well as over all observed events to reveal the impact of dropping misses.

It is important to point out that for other applications, there could be other ways to interrogate the timing errors that result from the cross-wavelet transform. Some of these possibilities are noted in the Discussion section.

3.2.4. Step 2d. Quantify Percent Hits Quantify the confidence in the timing error estimate

we report the hits to provide an assessment of the confidence in the timing error assessmentTo interpret our confidence in the timing error estimate, we can examine if the cluster maxs are the overlap between the significant areas of the observed WT and the significant areas of the XWT.

We can look at percent (%) overlap, that is, how many of the XWT events overlap with the WT events, either for all events or for the sampled event-sets. An overlap close to 0% would indicate that the model did not do a good job of simulating the observations – or it is a “miss” (flood is observed but not forecasted). If the overlap was 100%, it would be close to a perfect simulation.

Second, Iif we are looking at a single timing error for each event cluster, we may look to see if that event is significant in the XWT. If it is not, it gives us less confidence in the estimate. In Table 2, we can see for the prescribed 5 hour offset example, the cluster max was significant in the XWT. When there are multiple clusters for a given characteristic timescale, the % significance can be calculated as the ratio of the # of significant cluster max’s to the total number of cluster maxs.

We note that because we are calculating timing errors in terms of observed events, there is no information about “false alarms”, where a flood is forecasted but not observed.

4. Application of the Framework

The methodology developed in this paper is implemented in the R language and is made publicly available, as detailed in the code availability section at the end of the manuscript.

4.1. Data

The application of the methodology is illustrated using real and simulated stream discharge (streamflow, m3/s) data from four U.S. Geological Survey (USGS) stream gage locations: Onion Creek at US Highway 183, Austin, Texas (Onion Creek, TX; USGS site number 08159000), Taylor River at Taylor Park, Colorado (Taylor River, CO; USGS site number 09107000), Pemigewasset River at Woodstock, New Hampshire (Pemigewasset River, NH; USGS site number 01075000), and Bad River near Fort Pierre, South Dakota (Bad River, SD; USGS site number 06441500). We use the USGS instantaneous observations averaged on an hourly basis. NOAA's National Water Model (NWM, https://www.nco.ncep.noaa.gov/pmb/products/nwm/) is an operational model that produces hydrologic analyses and forecasts over the continental United States (CONUS) and Hawaii (as of version 2.0). The model is forced by downscaled atmospheric states and fluxes from NOAA's operational weather models. Next, the NoahMP (Niu et al 2011) land surface model calculates energy and water states and fluxes. Water fluxes propagate down the model chain through overland and subsurface (soil and aquifer representations) water routing schemes to reach a stream channel model. The NWM applies the three parameter Muskingum-Cunge river routing scheme to a modified version of the NHD-Plus version 2 (McKay et al. 2012) river network representation.

In this study, NWM simulations are taken from each version's retrospective runs (https://docs.opendata.aws/nwm-archive/readme.html). These are continuous simulations (not cycles) run for the period October 2010 to November 2016 and forced by the National Data Assimilation System (NLDAS)-2 product as atmospheric conditions. The nudging data assimilation was not applied in these runs either. We use NWM discharge simulations from versions V1.0, V1.1, and V1.2 (not all version may be publicly available).

To apply the methodology, we note that the observed and simulated datasets must be paired (overlapping). Further, for evaluation, any new simulation must also be paired with the observed. Missing data, which is common in observed time series, can be problematic and can result in false significance. We account for this our methodology by calculating the XT and XWT on each complete time series. This will be illustrated in the forthcoming example at Taylor River, CO.

4.2. Application

For illustration purposes we apply Steps 1 and 2 to an observed time series in Onion Creek, TX; for simplicity, we select an isolated peak (Figure 1a). First, we apply the wavelet transform to the observations (Figure 1b). This shows the time series in terms of its power by time and timescale, with warmer colors indicating more power. The black outline shows the areas of significance and the muted colors indicate the COI. To determine all observed events, we identify all the points that are significant and outside the COI (Figure 1c). Next, we average the power across each timescale: to the right of Figure 1b we show power averaged across all points for each timescale, and to the right of Figure 1c we show power averaged across just the events for each timescale. The latter is the one used to identify our characteristic scales. In this case, there is a single maximum at 22 hours. For the characteristic timescale, we see there is only 1 event cluster and the event with maximum power is marked with a star (Figure 1d).

For Step 2, we use the same Onion Creek, TX, peak from Figure 1a, and add a prescribed timing error of +5 hours to every point in the original time series (Figure 2a) to create a synthetic time series. We perform the cross-wavelet transform between the observed and synthetic time series (Figure 2b). The arrows in Figure 2b indicate the phase offset, which are used to calculate the timing error (Figure 2c). The timing error estimates show that for timescales greater than 10

~~hours, we get back the prescribed timing error of 5 hours, i.e., the scale must be at least double the timing error. In this case, because we are adding a prescribed error, the error is approximately 5 hours for all events, including for the characteristic timescale of 22 hours.~~

~~Finally, we repeat Step 2, but compare the observation of this event to actual model data from NWM V1.2. This shows that the model is early (Supplemental Figure 2a). We perform the cross wavelet transform (Supplemental Figure 2b) and examine the timing error (Supplemental Figure 2c). Table 1 summarizes the results: the mean error across the 22-hour characteristic timescale is -3.2 hours, as is the error for the cluster's maximum power. All events in the cluster are also significant in the XWT (100%), and the cluster maximum is also significant, providing confidence in this timing error estimation.~~

4~~5~~. Results

In the previous section, we illustrate the method using an isolated peak and a prescribed timing error. In this section, we ~~further~~ demonstrate the method~~, increasing the complexity by~~ using -NWM model ~~modeled~~ simulat~~ed data~~ions which introduce greater complexity -and longer time series ~~is used from several locations and time series to highlight the features of the method~~. Finally, we show ~~, finishing with~~ version-over-version comparisons for 5-year simulations to illustrate the utility for evaluation.

4.1 Demonstration using NWM data

~~5.1.~~ Pemigewasset River, NH

This example uses a three-month time series from the Pemigewasset River, NH, to ~~. First, we~~ examine ~~a three-month time series that exhibits~~ multiple peaks ~~above a base flow~~in the hydrograph (Figure ~~3~~4a). By eye, it is fairly straightforward to pick out three main peaks. From

Step 1 of our method, applying tThe wavelet transform onf the observations (Figure 34b and 34c), reveals up to three event clusters, depending on the characteristic timescale examined (Figure 43d). When we plot the average event power by timescale (right of Figure 34c), we see that there are nine relative maxima (small grey dots) – hence there are 9 characteristic scales for this example. -The -cluster maxima (grey stars) for each observed event cluster are shown in Figure 4d.

In Step 2, we compare the same observed time series from step 1 with output the simulation from the NWM V1.2 (Figure 45a): a), apply the cross-wavelet transform (Figure 45b colors), b) and calculate the timing error for all observed events (Figure 5b arrows), c) -subset the timing errors to the observed cluster maxima (Figure 5c stars), and d) retain only modeled hits (Figure 5c stars within the dashed contours). (Figure 45c). As previously mentioned, we are interested in the timing errors corresponding to observed events at the characteristic timescales. InTable 3 Figure 5a, the panels areis ordered by characteristic timescales from highest to lowest average power-; we only show the top 5 characteristic scales, using the first-subset of events, grouped by cluster. The first panel, where timescale = 24.8 hours, is tThe absolute maximum of the time average event spectrum has a timescale = 24.8 hours; (Figure 4c). This shows two cluster distributions: for cluster one, the model is lateclose to on-time (-0.052 hr) nearly 11 hours late and cluster two is lateearly (7hr-3.5 hours), both are hits, soand the average timing error is 3.5 hours late. However, for the next timescale (=27.8 hr), one of thethe third cluster maximums is a miss, so its the timing error is reported as a NA, and is not included in the average. This miss can be seen in Figure 5c where the last star falls just outside the XWT events. Moreover, this miss can also be interpreted from the comparison of the hydrographs in Figure 5a where the modeled third peak does not reasonably approximate the magnitude of the observed peak. Interestingly, the miss is a narrow miss at the shorter timescale of 27.8 hours while the associated (3rd) cluster maxima at the next most powerful characteristic timescale (33.1 hours) is a hit. This reflects that hydrograph is insufficiently peaked for this event but does have some of the observed, lower-frequency variability. Overall, this next most important characteristic timescale of 33.1 hours has timing results similar to the 27.8 hour timescale with the exception of the third cluster maximum. This raises the question if these are distinct characteristic timescales. In the Discussion and Conclusions section we discuss We point out that the for most events, and cluster two shows the model is early; the dark shading indicates that most of the events are significant in the XWT. The next two dominant scales of 27.8 and 33.1 hours have similar average power and are of the same order of magnitude at 27.8 hours and 33.1 hours; if we had applied smoothing to the graph of the time average event average power by timescale to address this issue, these relative maxima would smooth out. We will revisit this in the Discussion, when we discuss pathways to implementation in the Discussion and Conclusions.

The characteristic timescale with the next 4th highest time-average power maxima occurs at 111 hours, which is a different order of magnitude, suggesting that this may have a different physical process driving it. This At this timescale, the shows the model is to be late infor both event clusters (10 and 16 hours). , and rResults are similar for athe next timescale of 148 hours. We don't show results for the remaining 4 characteristic time scales with lower average power, since they have similar characteristic timescale values and associated timing errors to what has already been shown.

We can see how looking at the timing errors using the cluster distributions will get harder as the number of clusters increase, so it is also useful to summarize the information by looking at each cluster mean and max. If we run the methodology on the full 5-year Pemigewasset River time series, we can compare the mean and max timing errors for each characteristic time scale using box plots where the outline is shaded by the average confidence (Supplemental Figure 3). Table 2 summarizes this information. For example, the absolute maxima, at the 17.5 hour timescale has 86 clusters, and a timing error centered around zero (-0.43 hours), 75% of which are significant in the XWT. This is very similar to the results for the cluster max, as it is for the rest of the characteristic time scales. One other thing to note is that as expected, because the characteristic time scales are data driven, they are not the same as they were for the 3-month period.

5.2. Bad River, SD

The second example uses a two-month time series from the Bad River, SD, to illustrate the concept of consecutive peaks (Figure 6a). Whereas in the previous example it was fairly straightforward to pick out 3 distinct peaks, in this time series, there is one noticeable peak centered around June the 1st, with smaller peaks preceding and following it. The question is whether or not this is one event cluster or multiple? Looking at the wavelet transform (Figure 6b and 6c), we can see that for smaller timescales, there are more clusters, but for longer timescales, they are considered a single cluster.

In Step 2, we compare the same time series with output from NWM V1.2 (Figure 7a), calculate the cross-wavelet transform (Figure 7b), and calculate the timing error (Figure 7c). The timing error figure shows a sign switch: for longer timescales (i.e., when the peaks are considered part of a single event cluster), the model is early, but for shorter time scales (i.e., when the peaks are each considered their own cluster), the model is late. This is an important point: corrections at one scale may worsen timing error (or other metrics) at other scales.

This example has another interesting feature: namely that there is a false alarm in the model just before July 15. We note that because of our methodology, there is no observed event at that time, and therefore no timing error to be calculated, that is there is no information in the timing error statistics in terms of false alarms.

5.3. Taylor River, CO

In this example, we will examine a one-year time series from Taylor River, CO, that illustrates hydrograph peaks that are driven by different processes. The Taylor River is in a mountainous area where the spring hydrology is dominated by snowmelt runoff. To start, we will look at a portion of the spring melt season, where we can visibly see a diurnal signal (Figure 8). However, while it's easy to see that the model is too high in amplitude, it's hard to visually tell much about the timing error. Figure 9 shows that for the characteristic time scale of 23.4 hours, the model is usually early, with high confidence.

Supplemental Figure 4aFigure 67a shows athe year-long time series from Taylor River, CO, where we can see the snowmelt runoff in spring , butand also several peaks in summer, likely driven by summer rains. Supplemental Figure 4Figure 67b shows the WT, and also illustrates how missing data is handled: this results in additional COIs (muted colors) to account for the edge effects, and areas of the COI are ignored in our analyses.

From the statistically significant events in the WT, we again see the peak in the characteristic time scales at about 244 hours (right of Supplemental Figure 674c), but and there is another maxima at 99 and 118 hour timescales, relating to flows from the summer rains. This non-stationarity dominant timescale is evident in the wavelet power (Figure 6b and 6c). In Step 2, we compare the sameobserved time series with outputthe simulation from the NWM V1.2 (Figure 78a); here it is useful to zoom into the spring melt season time series (Figure 89), where we see that the amplitude of the diurnal signal is too high, but it's hard to visually tell much about the timing error. Next, the cross-wavelet transform (Figure 78b) and timing errors are calculated

(Figure 78c). The results are summarized in Table 4. ~~Looking at Figure 10, s~~Starting with the dominant 243 hour timescale, we see that ~~for the~~there are 11 clusters, that 73% (=8/11 cluster maxima) are hits, and that the model is ~~which are~~ generally early (the mean is 4.6 hours early)~~,~~ ~~and that 73% (=8/11 cluster maxs) are significant in the XWT.~~ ~~that are significant in the XWT, the model is generally early.~~ For the 118 and 99 hour timescale~~, the model is also early, b~~s, ~~ut none of the~~ ~~those cluster events~~maxs are ~~are not statistically significant in the XWT~~there are no hits. ~~(0%).~~ This suggests that we are confident in the ~~early~~ timing errors of the model for the diurnal snowmelt cycle, and these timing errors can ~~this could~~ be used as ~~qualitative~~ guidance for model performance and model improvements~~at this site until the model performance is improved~~. However, the model does not successfully reproduce key variability during the summer and timing errors are not valid at this timescale~~can not be used to evaluate or guide model improvements during this time.~~ ~~we show that it is less reliable for the early timing errors for the summer peaks.~~ This underscores the key point that timing errors are timescale dependent, and can help diagnose which processes to target for improvements.

~~Supplemental Figure 4b also illustrates how missing data is handled: this results in additional COIs (muted colors) to account for the edge effects, and areas of the COI are ignored in our analyses.~~

4.2 Evaluating Model Performance

Finally, we show how the methodology can be used for evaluating performance changes across NWM versions. We point out that none of the NWM version upgrades were targeting timing errors, so these results just provide a demonstration. We use ~~a~~ 5-year observed and modeled ~~overlapping time series and cluster max for the results~~time series at the three locations~~,~~:

Onion Creek, TX, and Pemigewasset River, NH, but cluster mean results were similar (not shown), and Taylor River, CO..

For Onion Creek, Table 5 summarizes the results for the three most important timescales and Figure 9 provides a graphical representation of these timing errors (hits only). For the NWM V1.0 for Onion Creek, we see that fFor the dominant 29.5 hour timescale and for all model versions, there were 197 cluster maximas, all of which were hits89.5% of which were hits, with a median timing error of 1.4 hours early, for which the median timing error is -1.4 hours, and all were significant in the XWT (Table 35). However, the model showeds progressively earlier timing errors with increasing version (Figure 9). The results are similar for the other two characteristic timescales.

Comparing V1.0, V1.1, and V1.2, the results for Onion Creek show that the median timing error has gotten slightly earlier (worse), although the distribution became tighter from V1.0 to V1.1 and V1.2 (Figure 1011). In Figure 1011, the dark blue color of the boxplot outline indicates that there is high confidence in the timing error, as the overlapping significance is close to 100% for the top three characteristic timescales. Using the 5-year overlapping time series forFor Pemigewasset River, NH, Table 6 summarizes the results for the 3 most important timescales and Figure 10 provides a graphical representation of the timing errors (hits only). At this location, we see that the median timing error improved by improved with NWM V1.2, getting closer to zero,. While the distribution of the timing errors became less biased than the previous versions, it also became but that the distribution became wider (Figure 1012). Over the timeseries, there were between 59 and 76 event clusters. Interestingly, the hit rate for all timescales was best for NWM V1.1 though its timing errors are broadly the worst. Again, the confidence is fairly highhits are fairly high (>80%) across characteristic time scales and versions

~~(Table 4~~6~~)~~From NWM V1.0 to NWM V1.2, improvements to both hit rate and median timing errors were obtained at all timescales.

~~, and >60 clusters were used in the estimations. Using 5-years from~~For Taylor River~~, CO (Supplemental~~ Table 7~~2, Supplemental Figure 5),~~ summarizes the results for the 2 most important timescales. ~~we see that f~~For the characteristic timescale of 235 hours (~10 days) there are only 4 event clusters~~, has low confidence~~but there are not and each model version has only 1 hit. The timing of this hit improves by roughly half its error from NWM V1.0 to NWM V1.2 in going from 16 to 9 hours. ~~many hits (~25%) for the 4 sampled clusters;~~ Tthe ~~timescale of~~ 23.4 hour timescale has 41 event clusters with a hit rate varying considerably by version. ~~s has a~~ The median timing error ~~that is~~ is fairly consistent with version, however, ranging from 6 to 7 hours ~~ly~~ early ~~by around 6 hours, with the version model confidence~~hits ~~ranging from 44% to 67% (Supplemental Table 2~~7~~)~~. ~~Results for the Bad River can be seen in Supplemental Table 3 and Supplemental Figure 6.~~

6. Discussion and Conclusions

In this paper, we develop a systematic, data-driven methodology to objectively identify timeseries (hydrograph) events and estimate timing errors in large-sample, high-resolution hydrologic models. The method was developed towards several intended uses: Primarily, it was developed for model evaluation, so that model performance can be documented in terms of defined standards. We illustrate this with the version-over-version NWM comparisons. Second, it can be used for model development, whereby potential timing error sources can be diagnosed (by timescale) and targeted for improvement. Related to this point, given the advantages of calibrating using multiple-criteria (e.g., Gupta et al. 1998), timing errors could be used as part of a larger calibration strategy. However, ~~as noted in the consecutive peaks example for the Bad River,~~ minimizing timing errors at one timescale may not translate to improvements in timing errors (or other metrics) at other timescales. ~~.~~ Wavelet analysis has also been used directly as an objective function for calibration, although a difficulty is in determining the similarity measure to use (e.g. Schaefli and Zehe 2009, Rathinasamy et al. 2014). Future research will investigate the ~~properties~~ application ~~of~~of the timing errors presented here for calibration purposes. Finally, the approach can be used for model interpretation and forecast guidance~~,~~ as estimating timing errors provides ~~a~~ characterization of the timing uncertainty (i.e., for a given timescale, the model is generally late or early)~~, as well as a measure of the~~ or confidence~~, that could be useful for qualitative forecast guidance~~.

Given the fact that several subjective choices were made specific to our application and goals, ~~we think~~ it is important to highlight that we have made the analysis framework openly available (detailed in the code availability section below), so the method can be adapted, extended, or refined by the community right away. ~~For instance, because of our focus on model evaluation and development, we use the observed WT to identify events. However, i~~In other ~~instances~~ applications ~~it might be sufficient to only sample events that are in the significant areas of the XWT (essentially to identify the characteristic scales and event-set directly from the XWT instead of from the WT). This might be reasonable for applications that are more focused on model interpretation in a real-time forecasting mode, but it would not allow for version comparison and it is not guaranteed that all the important characteristic scales would be identified (i.e., the model may not capture some real-world processes, and therefore miss the associated characteristic timescales).~~ We ~~only~~ look at ~~the~~ timing errors from an observed event-set relevant to our analysis, but there are other ways to subset the events that might be more suitable to other applications. For instanceexample, we focus on the event cluster maxima, but one could also examine the event cluster means or the local maxima along time. AlsoAnother alternative to, instead of finding the event of maximum power in eachevent cluster maxima (i.e., for a given timescale), it would be possible to identify the event with maximum power in “islands of significance” across timescales, i.e., contiguous regions of significant areas contiguous significance in time across both time and timescales. However, tThis approach would ignore s that multiple frequencies can be important at once. Also and defining the such islands is also not straightforwardwhen connected is problematic. Yet anotherA differenceet approach could be desirable Iif one suspected non-stationarity in the characteristic timescales over the timeseriestime series,. tThen perhaps a different approach such as a moving average in timescale could be employed to identify characteristic timescales. For instance, Further, In our approach, we define the event set broadly,. However, but it could be subset for high peak or using streamflow thresholds (e.g. for flooding events) to compare events in the wavelet domain with traditional peak-over-threshold approachesevents. For example, Supplemental Figure 117 shows the maximum streamflows for the event-set from the 5 year run at Taylor River. This figure shows that all events identified by the algorithm are not necessarily high flow events (i.e., the maximum streamflow peaks are lower for the 23.4 hour timescale as compared to the 235.6 hour timescale). To compare with traditional peak-over-threshold approaches, this event-set could be filtered to include only events above a given threshold (i.e., events in both the wavelet and time domains). ; this event-set could be filtered to include only events above a given streamflow threshold (i.e. events in both the wavelet and time domains). The method provides a quantification of the confidence in thepercent hits for the timing errors; however, , and we include all timing errors in our summaries, whether they are hits or misses. However, iIt might make more sense to drop those points in the timing error assessment, that do not have a high confidence (i.e., with a low percent of events that significantly overlap between the XT and the XWT) and to and to only calculate the timing errors on hits.flag those events as misses.

Another point that arises is how many characteristic timescales should be examined and the similarity of adjacent characteristic timescales. HereIn our method, we average the power across in timescales and identify characteristic scales to be at every absolute and relative maxima. As seen in the illustrative examples, this can result in multiple characteristic scales, some of which can be quite similar, suggesting that events at those scales are from similar or related processes. One solution could be to smooth the average power by timescale, which would reduce the number of local maxima, or to look at timing errors within a band of timescales. It is also important to note that the characteristic scales are data-driven, so they will change with different lengths of observed time series. Longer runs capture more events and should converge on the more dominant timescales and events for a location. However, for performance evaluation, overlapping time periods for observed and modeled time series are needed.

In our application of the WT, we follow Liu et al. (2011) and select the Morlet as the mother wavelet. However, results are sensitive to the mother wavelet selected. Further discussion of mother wavelet choices can be found in Torrence and Compo (1998) and in ElSaadani and Krajewski (2017).

In shortsummary, this paper provides a systematic, flexible, and computationally efficient methodology for calculating model timing errors that is appropriate for model evaluation and comparison, and is useful for model development and guidance. Based on the wavelet transform, the method introduces timescale as a property of timing errors. The approach also identifies streamflow events in the observed and modeled timeseries and only evaluates timing errors for modeled events which are hits in a 2-way contingency analysis. Future work will apply the approach to identify characteristic timescales across the United States, as well as to assess the associated timing errors in the NWM.

Code/Data Availability

The code for reproducing the figures and tables in this paper are provided in the public github repository https://github.com/NCAR/wavelet_timing with instructions for installing dependencies. The core code used in the above repository is provided in the public "rwrfhydro" R package https://github.com/NCAR/rwrfhydro. The code is written in the open-source R language (R Core Team 2019) and builds off multiple, existing R packages. Most notably the wavelet and cross-wavelet analyses are performed using the "biwavelet" package (Gouhier et al. 2018).

We emphasize that the analysis framework is meant to be flexible and adapted to similar applications where different statistics may be desired. The figures created are specific to the applications in this paper but provide a starting point for other work.

~~The code for reproducing the figures in this paper as well as extended vignettes/notebooks are provided in public github repository https://github.com/NCAR/wavelet_timing. In addition to reproducing the analyses and figures in this paper, several jupyter notebooks provide more detailed analyses of the time series included in this paper. We emphasize that the analysis framework is meant to be flexible and adapted to similar applications where different statistics may be desired. The figures created are specific to the applications in this paper but provide a starting point for other work.~~

~~The core code is provided in the public "rwrfhydro" R package https://github.com/NCAR/rwrfhydro. The package can be installed as described by the README document in the repository and in the Supplemental Online Materials for this paper. The code is written in the open-source R language (R Core Team 2019) and builds off multiple, existing R packages. Most notably the wavelet and cross-wavelet analyses are performed using the "biwavelet" package (Gouhier et al. 2018).~~

Credit Author Statement

ET and JLM collaborated to develop the methodology. ET led the results analysis and manuscript preparation and revisions. JLM developed the initial idea for the work, the open source software, and visualizations.

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

Acknowledgements:

The authors would like to thank Dave Gochis for useful discussions and Aubrey Dugger for providing NWM data. We thank the NOAA/OWP and NCAR NWM team for its support of this research. This research is funded by the NOAA Office of Water Prediction and the Joint Technology Transfer Initiative grant 2018-0303-1556911. This material is based upon work supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977.

References:

Bogner, K., Pappenberger, F., 2011. Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. Water Resour. Res. 47, W07524, doi:10.1029/2010WR009137.

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values, Springer Ser. Stat. Springer, London.

Daubechies, I., 1990. The wavelet transform time-frequency localization and signal analysis. IEEE Trans. Inform. Theory 36, 961-1004.

ElSaadani, M., Krajewski, W. F., 2017. A time-based framework for evaluating hydrologic routing methodologies using wavelet transform. *Journal of Water Resource and Protection*, *9*(7), 723–744.

Ehret, U., Zehe, E., 2011. Series distance-an intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. *Hydrology and Earth System Sciences*, *15*, 877–896. https://doi.org/10.5194/hess-15-877-2011

Gochis, D.J., M. Barlage, R. Cabell, M. Casali, A. Dugger, K. FitzGerald, M. McAllister, J. McCreight, A. RafieeiNasab, L. Read, K. Sampson, D. Yates, Y. Zhang (2020). The WRF-Hydro® modeling system technical description, (Version 5.1.1). *NCAR Technical Note.* 107 pages. Available online at: https://ral.ucar.edu/sites/default/files/public/WRFHydroV511TechnicalDescription.pdf. Source Code DOI:10.5281/zenodo.3625238

Gouhier, T.C., Grinsted, A., Simko, V., 2018. R package biwavelet: Conduct Univariate and Bivariate Wavelet Analyses (Version 0.20.17), https://github.com/tgouhier/biwavelet.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G. F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology, 377(1-2), 80-91.

Gupta, H.V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andréassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth. Hydrol. Earth Syst. Sci. 18, 463-477, https://doi.org/10.5194/hess-18-463-2014.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Towards improved calibration of hydrologic models: multiple and non-commensurable measures of information. Water Resources Research 34(4): 751–763.

Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. *Hydrological Processes*, *22*(March), 3802–3813. https://doi.org/10.1002/hyp

Koskelo, A. I., Fisher, T. R., Utz, R. M., and Jordan, T. E., 2012. A new precipitation-based method of baseflow separation and event identification for small watersheds (< 50 km2), J. Hydrol., 450–451, 267–278, https://doi.org/10.1016/j.jhydrol.2012.04.055.

Lane, S.N., 2007. Assessment of rainfall–runoff models based upon wavelet analysis. Hydrol. Processes 21, 586–607, https://doi.org/10.1002/hyp.6249.

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

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

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

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

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

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

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

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

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

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

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

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

Seibert, S. P., Ehret, U., Zehe, E. 2016. Disentangling timing and amplitude errors in streamflow simulations. Hydrology and Earth System Sciences, 20, 3745–3763. https://doi.org/10.5194/hess-20-3745-2016

Torrence, C., Compo, G.P., 1998. A Practical Guide to Wavelet Analysis. Bull. Am. Meteorol. Soc. 79(1), 61-78.

Veleda, D., Montagne, R., Araujo, M., 2012. Cross-wavelet bias corrected by normalizing scales. J. Atmos. Ocean. Technol. 29, 1401-1408.

Weedon, G.P., Prudhomme, C., Crooks, S., Ellis, R.J., Folwell, S.S., Best, M.J., 2015. Evaluating the performance of hydrological models via cross-spectral analysis: case study of the Thames Basin, United Kingdom. J. Hydrometeorol. 16(1), 214–231. http://doi.org/10.1175/JHM-D-14-0021.1.