

~~This information product has been peer reviewed and approved for publication as a preprint by the U.S. Geological Survey.~~

Technical note: A low-cost approach to monitoring relative streamflow dynamics in small, headwater streams using timelapse imagery and a deep learning model

Phillip J Goodling¹, Jennifer H Fair², Amrita Gupta³, Jeffrey D Walker⁴, Todd Dubreuil², Michael Hayden², Benjamin H Letcher²

¹U.S. Geological Survey, Earth System Processes Division, 5522 Research Park Drive, Catonsville, Maryland, 21228

²U.S. Geological Survey, Eastern Ecological Science Center, S. O. Conte Research Laboratory, One Migratory Way, Turners Falls, MA 01376

³Microsoft Corporation AI For Good Lab

⁴Walker Environmental Research, LLC

Correspondence to: Phillip Goodling (pgoodling@usgs.gov)

Abstract. Despite their ubiquity and importance as freshwater habitat, small headwater streams are under monitored by existing stream gage networks. To address this gap, we describe a low-cost, non-contact, and low-effort method that enables organizations to monitor relative streamflow dynamics in small headwater streams. The method uses a camera to capture repeat images of the stream from a fixed position. A person then annotates pairs of images, in each case indicating which image has more apparent streamflow or indicating equal flow if no difference is discernible. A deep learning modelling framework called Streamflow Rank Estimation (SRE) is then trained on the annotated image pairs and applied to rank all images from highest to lowest apparent streamflow. From this result a relative hydrograph can be derived. We found that our modelled relative hydrograph dynamics matched the observed hydrograph dynamics well for 11 cameras at 8 streamflow sites in western Massachusetts. Higher performance was observed during the annotation period (median Kendall's Tau rank correlation 0.75 with range 0.6-0.83) than after it (median Kendall's Tau 0.59 with range 0.34 – 0.74). We found that annotation performance was generally consistent across the eleven camera sites and two individual annotators and was positively correlated with streamflow variability at a site. A scaling simulation determined that model performance improvements were limited after 1,000 annotation pairs. Our model's estimates of relative flow, while not equivalent to absolute flow, may still be useful for many applications, such as ecological modelling and calculating event-based hydrological statistics (e.g., the number of out-of-bank floods). We anticipate this method will be a valuable tool to extend existing stream monitoring networks and provide new insights on dynamic headwater systems.

35 **1 Introduction**

Small headwater streams make up 50-70% of stream network length (Benda et al., 2004; McManamay and DeRolph, 2019) and are fundamental units of riverine networks. Streamflow dynamics in these streams are crucial controls on aquatic ecosystem function (Carlisle et al., 2017; Colvin et al., 2019; Hitt et al., 2022), thermal changes, and the routing of sediment and contaminants. Headwater streamflow dynamics are uniquely complex for the following reasons: 1) a majority of small
40 (second-order or less) stream channels dry out seasonally or during drought events (Jaeger et al., 2021; Messenger et al., 2021), 2) along-channel changes can be abrupt due to geologic controls and focused groundwater inputs (Briggs et al., 2018), and 3) due to small catchment size, these streams are particularly susceptible to drastic hydrologic alterations, both anthropogenic (damming, impervious surface runoff) and natural (ice or beaver damming, wildfire effects, geomorphic changes).

Despite their importance and vulnerability, headwater and non-perennial streams are underrepresented by streamflow
45 monitoring networks in the United States. (Deweber et al., 2014; Seybold et al., 2023) and across the world (Krabbenhof et al., 2022). Three primary limitations lead to a sparse headwater monitoring network: first, monitoring and maintaining traditional stage-discharge gage records (Turnipseed and Sauer, 2010) to a high quality requires expertise and training that limits the number of organizations able to collect the records. Second, velocity measurements in small, shallow, and slow-moving streams are difficult to collect and have high uncertainty, making the percentage error of streamflow discharge much
50 higher in small streams than large streams (Horner et al., 2018; King et al., 2022; Levin et al., 2023; McMillan et al., 2012). Third, in-stream instruments to measure stage in headwater streams are frequently lost or damaged due to shifting streambeds, very high local velocities, and beaver or other animal activity. Even disregarding the challenges in collecting the data, where streams are non-perennial or form disconnected pools, traditional pressure transducer-based stage measurements provide incomplete information regarding (dis)connectedness of the stream channel, making these records inadequate for certain uses
55 in ecohydrological modelling (Steward et al., 2012).

Streamflow monitoring using imagery is an attractive alternative to in-stream instruments and has grown in popularity as camera technology has improved. Collecting imagery is appealing because it requires very little training or specialized equipment. However, analysing a large volume of imagery can be a challenge; a range of approaches has been introduced to date. Initially, manual interpretation (Schoener, 2018) or rules-based image processing techniques (Chapman et al., 2022;
60 Gilmore et al., 2013; Leduc et al., 2018; Noto et al., 2022) were used to automate the reading of a staff gage placed in the channel. While effective and low-cost, these staff-plate based approaches still require the installation of in-channel infrastructure that may not be permitted in protected lands or can be damaged by high flows. Additionally, stage monitoring is restricted to the location of the staff plate; therefore, any debris on the staff plate or view blockage due to snow or vegetation will result in missed readings. Computer-vision based approaches that avoid the use of an in-channel staff plate have been
65 introduced, but generally require the manual identification of a specific region of interest in the image (Keys et al., 2016), image orthorectification using ground control points, and detailed high-resolution 3D models of riverbed and bank geometry to estimate changes in stage (Eltner et al., 2018).

Advances in deep learning approaches for imagery analysis have created new opportunities for environmental monitoring. For example, several recent studies have applied deep learning to image-based stream stage monitoring to eliminate the need for fixed in-stream staff plates. Many of these papers use established image segmentation algorithms (i.e. convolutional neural networks) to classify parts of the image as “water” or “not-water” (Eltner et al., 2021; Liu and Huang, 2024; Vandaele et al., 2021). Using a reference point on the image and knowledge of the interface location, the stream level is tracked over time. While effective, these approaches are sensitive to channel rearrangement or view blockage at the water/not-water interface. They also still require some manual judgement about the location of interest in the image frame for which stage is provided and image orthorectification using ground control points.

Unlike other deep learning approaches for streamflow estimation, Streamflow Rank Estimation (SRE) was developed to minimize the need for external monitoring data to train a model (Gupta et al., 2022). The approach aims to estimate streamflow dynamics without the need for traditional discharge observations, an in-channel staff plate, designating a region of interest, or imagery orthorectification. SRE uses a learning-to-rank framework that is trained using many pairs of stream images, with discharge in the images of each pair visually compared, removing the need for stream discharge training data. We refer to the person-generated pairwise ranks as “annotations”. The model is trained using the annotations to sort images from high apparent streamflow to low apparent streamflow by fine-tuning a convolutional neural network (a ResNet-18 (He et al., 2015) architecture pretrained on ImageNet (Deng et al., 2009)) and using a learning-to-rank approach utilizing the RankNet loss function (Burgess et al., 2005). The rank of each image can be used to create a streamflow percentile which is correlated with the streamflow discharge and can be interpreted as a dimensionless hydrograph. While the absolute streamflow could be estimated from the streamflow percentile using an assumed streamflow discharge distribution, for unmonitored catchments this distribution would need to be estimated independently of the SRE model and would be a significant source of uncertainty in absolute streamflow estimates (Gupta et al., 2022). As a trade-off for low-effort model training and minimal external information requirements, the rank-based streamflow percentile estimate is the primary output produced by the SRE model.

To date, the SRE model has been tested at a limited number of sites with simulated annotations derived from known streamflow discharge timeseries, but not with annotations created by people. With simulated annotations, SRE characterized streamflow percentile dynamics with a Kendall’s rank correlation greater than 0.7 in five of six stream locations (Gupta et al., 2022). The number of annotations (n = 500, 1000, 2500, 10000) and annotators’ ranking ability (could discern 0%, 10%, 20%, 50% discharge difference) both strongly influenced the model’s ranking performance. This promising early work motivated us to further evaluate the real-world performance of the model by using person-generated annotations and expanding the number of stream sites at which we assessed model performance. With a better understanding of the factors influencing model performance, we plan to apply SRE to currently unmonitored headwater catchments.

This paper describes a methodology for monitoring [relative](#) streamflow dynamics in small headwater streams using timelapse imagery coupled with a deep learning model trained using person-generated annotation. We evaluate the real-world performance of this monitoring system and answer the following questions:

1. How accurate are people at ranking images by streamflow?
2. How accurate are the image-derived relative hydrographs developed using person-generated annotations?
3. Which factors influence ranking model accuracy and can indicate which unmonitored catchments would be suitable for low-cost camera monitoring?
4. How many person-generated annotations are required to achieve stable ranking model performance?

2 Methods

2.1 Data Collection

To collect timelapse imagery from low-cost cameras, this project developed a web platform titled the Flow Photo Explorer (<https://www.usgs.gov/apps/ecosheds/fpe/>). Since its inception in October 2021, the Flow Photo Explorer (FPE) platform has accepted imagery submissions from an array of organizations with a common motivation of enhancing and expanding stream monitoring networks. While guidelines are provided on the webpage, there are few restrictions on how cameras are configured and what views they capture. The only requirement is that the imagery format uploaded to the FPE platform is formatted with EXIF metadata, which is a common imagery data format across many low-cost battery-powered game or trail cameras. We recommend a photo every 15 minutes, though the FPE database contains intervals from less than 5 minutes to once per day. The recommended camera view is looking downstream or upstream, though based on field conditions some sites may instead feature cross-stream or tangential views. We expect that the image-based monitoring approach will work best when at least some fixed objects (i.e. trees, boulders, bridge pilings, stream banks) are visible at all levels of streamflow. An example camera view with these fixed features visible is shown in Fig. 1. If a user knows a nearby U.S. Geological Survey (USGS) stream gage [monitoring the same stream reach](#), they can indicate the USGS station identifier and data are automatically pulled from the USGS National Water Information System (U.S. Geological Survey, 2024) database. Alternatively, they can upload their own streamflow observations, although they are not required. To test the methodology, we co-located 11 cameras with eight USGS gages in western Massachusetts for which records of stream discharge are available (Fair et al., 2025). Four cameras were located at the same streamflow monitoring location to examine the effect of differing camera angles on monitoring performance. In this study we collected imagery every 15 minutes with Reconyx (Hyperfire 2 model) Bushnell (Trophy and Essential models) cameras that were mounted to trees (except for one site that was affixed to a bridge) using swivel mounts and a secure metal housing.



Figure 1 – The recommended camera view includes stream banks and fixed objects such as trees or boulders visible at most flows. Photograph by the U.S. Geological Survey.

For this analysis we set minimum data availability criteria to test the method at sites with sufficient data. We expected that seasonal changes in vegetation, streamflow, and snow cover would appear in the imagery. Therefore, we selected sites with stream discharge and imagery data that spanned at least 1.5 years. We implemented this criterion to ensure that the model training period spanned at least one full year, so that all seasons were represented, and so that we additionally had access to a final half year of data for testing purposes. Within this span, we allowed some data gaps, since these are common in our available set of imagery data. We required at least 180 complete days of data within the 1.5 years, which is a completeness of approximately 33%. **Table 1** contains a list of sites that met our data availability requirements. These locations are mapped in **Fig. 2**. In this analysis we used daytime-only imagery (from 7 am to 7 pm), though many sites have cameras with an infrared flash that also produce usable imagery at night.

To guide user site selection and setup, we evaluated patterns in model performance according to two key site attributes. The first is a measure of flow variability during the monitoring period. Some streams, such as those heavily influenced by groundwater discharge can have small fluctuations in stream stage that are difficult to identify in imagery. We selected the coefficient of variation (CV) of log-transformed streamflow $\log_{10}(Q)$ to quantify the general variability of the stream. The second metric is a simple qualitative assessment of how stable the camera view is over the period of record. This metric is primarily for quantifying if there were abrupt changes in the field of view of the image time series, mainly coinciding with when the camera was serviced. Cameras can also shift slightly due to vibrations or wind changing the mounting position, though these types of shifts are minor alterations compared to abrupt view changes. In this rating system, a camera stability value of “Low” indicates that there was at least one camera view change of 50% or greater (i.e. only half of the original frame was still visible). “Medium” indicates at least one camera view change between 25% and 50%, while “High” indicates that all view changes were below 25%. These two attributes were selected to inform user site selection and field methods.

Location ID	Station Name (USGS Station ID)	Monitoring Period	% of images have observed stream flow	Number of Annotations	Training period CV of $\log_{10}(Q)$	Camera stability	Drainage area (km ²)
ABB	Avery Brook Bridge (01171000)	2021-03-10 to 2024-04-02	99.1	3,147	0.8	Low	7.8
ABL	Avery Brook River Left (01171000)	2021-07-02 to 2024-04-02	98.8	2,277	0.8	High	7.8
ABR	Avery Brook River Right (01171000)	2021-03-19 to 2024-04-02	99.3	2,214	0.8	Medium	7.8
ABS	Avery Brook Side (01171000)	2021-03-19 to 2024-04-02	99.2	2,441	1.0	High	7.8
GR	Green River (01170100)	2022-09-29 to 2024-03-29	99.2	5,057	0.2	High	107.2
SB	Sanderson Brook (01171010)	2021-04-01 to 2024-03-22	70.9	4,821	1.2	Low	4.4
WB0	West Brook 0 (01171100)	2022-02-01 to 2024-04-02	99.1	7,953	0.8	High	27.7
WBL	West Brook Lower (01171070)	2019-02-27 to 2024-04-09	67.7	2,256	0.7	High	21.8
WBR	West Brook Reservoir (01171020)	2021-03-25 to 2024-03-22	64.8	2,325	1.1	High	16.1
WBSR	West Branch Swift River (01174565)	2017-09-14 to 2024-03-28	99.5	3,553	0.3	Medium	32.6
WW	West Whately (1171005)	2021-04-06 to 2024-04-09	70.2	2,510	-2.5	Medium	1.3

Formatted Table

Formatted Table

Formatted Table

Formatted Table

Formatted Table

Formatted Table

Table 1 - Summary of data collected at locations included in this analysis. Streamflow observations were originally reported in a [U.S. Geological Survey \(USGS\) data release](#) (Fair et al., 2025). "Training period CV of $\log_{10}(Q)$ " refers to the coefficient of variation of log-transformed streamflow discharge during the model training period.

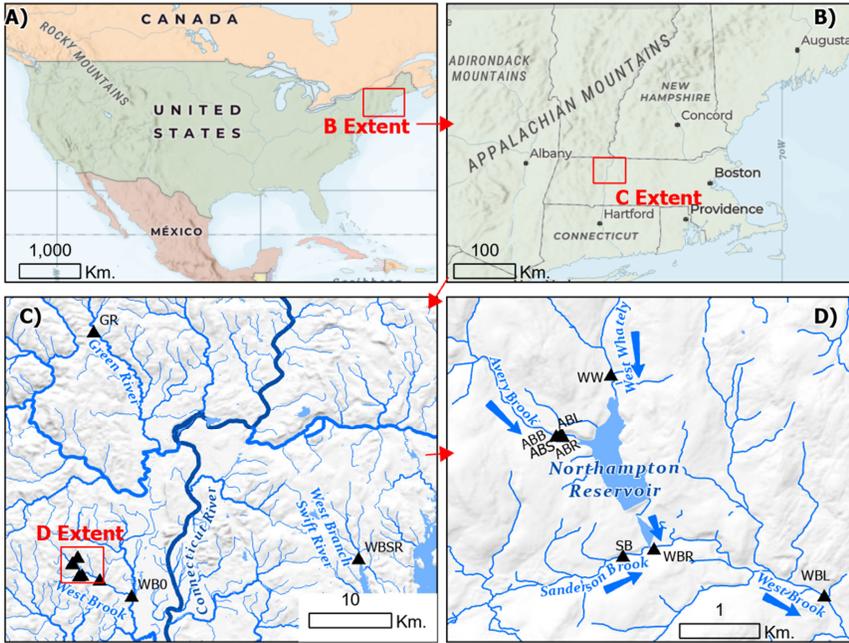


Figure 2 - Map of monitoring locations in western Massachusetts, USA (Fair et al., 2025; Goodling et al., 2025). Triangles in panels C and D indicate monitoring sites and are labelled with site identifiers listed in Table 1. Arrows in Panel D indicate streamflow direction. Water bodies shown are from the NHDPlus Version 2 (McKay et al., 2012) (panel C) and NHD High Resolution (Moore et al., 2019) (panel D) datasets.

2.2 Data Annotation

Training the neural network model to predict streamflow dynamics from imagery requires external site-specific information. Because we hope to use this method in places with no other information except for the imagery, we could not use any streamflow data in model training. Instead, we relied on people to rank pairs of images by streamflow in a process called ‘data annotation’. In the FPE web application, users were shown two photos from a given site side-by-side and asked to indicate which one had more streamflow (Fig. 3-3). The images selected to form a pair were selected at random. The users also indicated if the images appear “about the same” or if the image was a “bad photo” (obscured or too dark). “Don’t know” was selected if the photo is bad or if other aspects of the images made them difficult to compare, such as a large difference in camera view or camera angle. Image pairs marked “don’t know” were not used in model training. In this study, users were only presented with images collected during daytime (7am – 7pm). A typical user completed an annotation in 1-3 seconds on average; if focused, an individual could perform approximately 1000 annotations in an hour. Our dataset includes 17 unique

Formatted: Font: Italic

Formatted: Font: Italic

Formatted: Font: Italic

annotators who contributed to the model training; however, only two annotators represent 93.7% of all the annotations and we focus on these two in our discussion of annotator performance. Both of these annotators were student interns (one ecology graduate student, one environmental science undergraduate student). The student interns were associated with the project but had no specialized training or experience in streamflow monitoring.

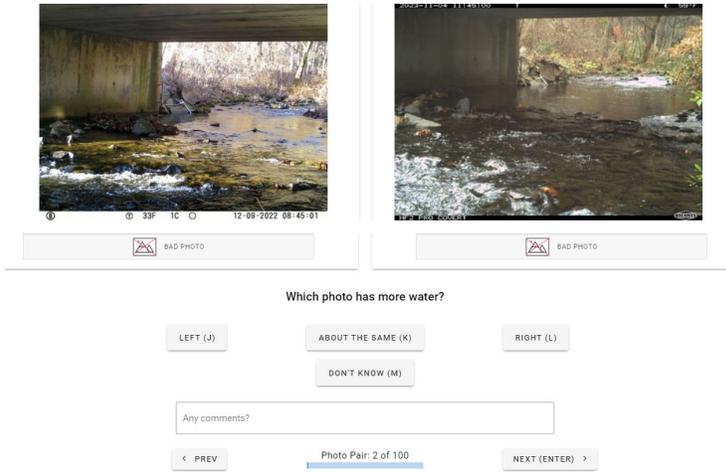


Figure 3 – The web-based annotation interface from the Flow Photo Explorer used in this study to develop training datasets for the ranking model.

The process of annotation was not error-free; the judgments made by individual annotators could sometimes be incorrect. This could be through simple errors of transcription (i.e. clicking the incorrect button) or because the imagery pairs were difficult to compare because of lighting, vegetation, or seasonal differences. These errors, if significant, could provide spurious information to the deep learning model. We therefore quantified the performance of our annotation dataset using the known true flow-based ranks from the co-located USGS gage data. Our primary metric was classification accuracy for the selection of the “left” or “right” image with higher streamflow in the image pair:

$$(1) \quad \text{Classification \% Accuracy} = \frac{TL+TR}{TL+FL+TR+FR} * 100$$

Where TL and TR refer to true left and true right selections and FL and FR refer to false true and right selections. We observed that the difficulty of the selection increases, and therefore the classification accuracy decreases, if the two photos had similar streamflow. To fully describe annotation performance, we provide our metrics as functions of the relative flow difference between the images. The relative flow difference (Δ_{rel}) between a pair of photos shown to an annotator was calculated as:

$$(2) \quad \Delta_{rel} = \frac{|Q_1 - Q_2|}{\frac{1}{2}(Q_1 + Q_2)}$$

Where Q_1 and Q_2 represent streamflow values for the two images. For positive values inclusive of zero, the value Δ_{rel} is bounded to be between zero and two. A Δ_{rel} value near zero indicates close agreement between Q_1 and Q_2 whereas a Δ_{rel} value of 2 could indicate that one of the two values is approaching zero or infinity. We compute the overall classification accuracy within binned increments of $0.1 \Delta_{rel}$; the unweighted binned performance is used to develop a function describing the relationship between Δ_{rel} and classification accuracy.

2.3 Modelling Methodology

Annotated images were ranked into an ordered sequence using the previously developed SRE neural network model (Gupta et al., 2022). ~~An independent model was trained for each site.~~ The SRE neural network model takes an image as input, which includes three channels (RGB), and generates a dimensionless, continuous-valued score representing relative streamflow as output. The score is derived by applying a sequence of mathematical operations to the input image, including spatial convolutions, which help the model extract relevant features from the image. During training, the model is given batches of paired images ranked by annotators based on relative streamflow. ~~The training objective is Two neural networks with shared model weights sequentially predict dimensionless scores for the model two images. The pair of scores is used to compute a probabilistic ranking loss (Burges et al., 2005) learn to assign that is minimized when the model assigns~~ a higher score to the image that the annotator ranks as having higher flow, or ~~to assign~~ assigns the same score to both images if the annotator ranks them as having the same flow. ~~This architecture is sometimes called a “twin neural network”.~~ Images are pre-processed by resizing, centre-cropping to exclude metadata bands, and normalizing. While training, data augmentations such as random crops, horizontal flips, rotations, and colour jitter are applied to improve model robustness, generalization, and reduce overfitting (Shorten and Khoshgoftaar, 2019). Additional detail on ~~model development and~~ image pre-processing is available in the **supplemental materials**. After training, the model is used to generate score predictions for all images from a site, which are then standardized into z-scores by subtracting the mean and dividing by the standard deviation.

The imagery data were divided into training, testing, and validation splits to enable robust model evaluation. Unlike many machine learning applications, the model learns from image pairs and not individual images; therefore, these splits are a bit more complex to develop. When reporting model performance, we identify images that comprised pairs used for training (“train²³”, representing 80% of annotations) or validation (“val²³”, representing 20% of annotations). Images that were not part of any annotation pair provided to the model are used for “test”. We further divided this into “test-in”, which is coincident with the timeframe of annotation, and “test-out” (when available) for the period following the period with annotations. “All-in” is the combined set of images, regardless of if they are part of an annotation pair, during the annotation period. “All” is the performance for all images. We consider “test-in” to represent a retrospective model performance, while “test-out” to represent the expected performance of a deployed operational model on new imagery.

The sites in this study were co-located with traditional USGS streamflow gages, which enables us to evaluate model performance relative to these instruments. Our model performance metric is Kendall’s Tau, a nonparametric rank-based correlation coefficient (Kendall, 1938). We selected this metric because it is insensitive to monotonic transformations such as

225 log-transformation and percentile calculations, making it appropriate to compare values on different scales and with different
distributions. As a metric it is strict regarding timing; short-lived peaks, if slightly mis-timed, will result in low Kendall's Tau.
Because it is based on ranks, it is insensitive to the magnitude difference between two values. As a result, low-flow
observations, which are more common, have a greater influence on the resulting Kendall's Tau than short-lived high-flow
observations.

230 To provide a preliminary understanding of the factors influencing model performance we present pairwise
relationships between annotation accuracy, streamflow variability, camera stability, and model performance. For comparisons
among the numeric values we present the Pearson's correlation coefficient and two-sided p-value calculated with the *cor.test*
function in R version 4.3.2 (R Core Team, 2021). For comparisons between numeric values and the categorical camera stability
metric, we present the results of the nonparametric Kruskal-Wallis test to evaluate if the distribution varies among the
235 categories (Kruskal and Wallis, 1952). If significant, we perform Dunne's *post hoc* pairwise multiple comparison test to
identify which categories have statistically different distributions (Dunn, 1964). The Kruskal-Wallis and Dunne's tests are
computed with the *rstatix* R package (Kassambara, 2023).

2.4 Sensitivity analysis

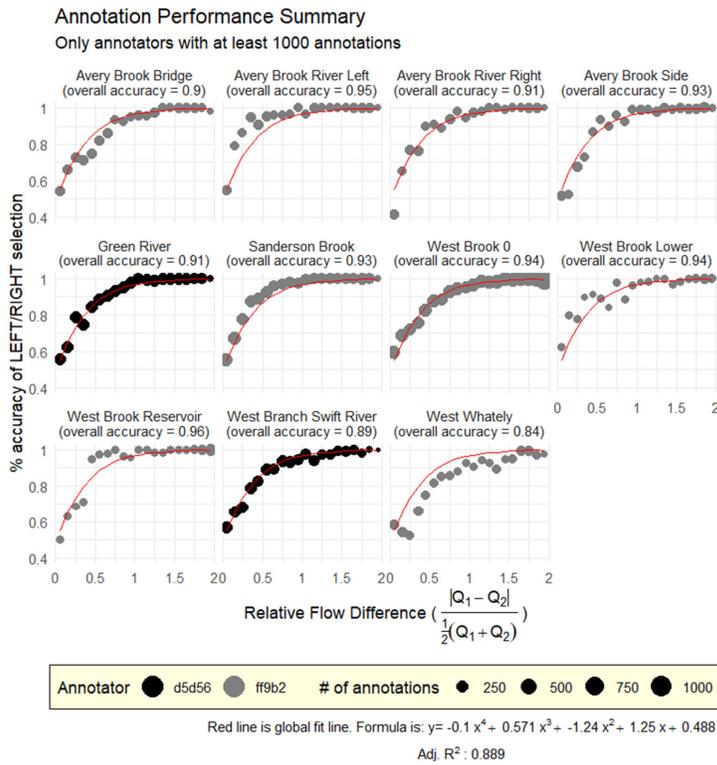
We performed a sensitivity analysis to understand how many person-generated annotations are required to achieve
240 acceptable performance. In this case, the target performance level was that achieved by training the model with all available
image pair annotations for a given site. We created nested subsets of the annotations, beginning with increments of 100 up to
500, then using larger increments of 250 up to 1500, and finally using increments of 500 up to 3000, with additional subsets
at 4000 and the maximum number of available annotations. Smaller increments were used at the lower end of the annotation
range to capture the more substantial improvements in model performance that are typically observed with initial increases in
245 training data. Each subset was a strict superset of the previous one, meaning that each larger subset contained all the pairs from
the smaller subsets plus additional pairs. This allowed us to assess how increasing the volume of training data impacts model
performance and to identify the point where performance plateaus, avoiding unnecessary annotation efforts that may not
significantly improve performance. The sensitivity analysis reported the Kendall's Tau model performance metric is for the
"test-in" data split for daytime images (7am – 7pm).

250 To ensure the robustness of our findings, the analysis was repeated five times. For each repetition, we randomly
permuted the order of the annotations before generating the nested subsets, thereby mitigating any potential variance that could
arise from the specific sequence of training samples.

3 Results

3.1 Annotation results

255 Annotation performance in our dataset was high (average 92.2% accuracy) and was generally consistent across sites
and annotators. Accuracy was well-described by an increasing function of the relative flow difference (global 4th order
polynomial, $R^2 = 0.89$, **Fig. 4**, red lines). At all sites, annotation accuracy neared 100% accuracy above a relative flow
difference of 1 (which occurs when one image has three times as much streamflow as the other). As the relative flow difference
neared 0, classification accuracy approached 50%, which is equivalent to guessing between the photos. Similar curves are
260 observed for the two primary annotators (represented by symbols in **Fig. 4**). To characterize the overall accuracy of the
annotation at a site, the percent accuracy of all annotations regardless of relative flow difference is reported in each panel of
Fig. 4. The site with the lowest overall annotation performance—West Whately, with an 84% overall accuracy—had the lowest
streamflow coefficient of variation a “medium” level of camera stability (**Table 1**).



265

Figure 4 – Annotation accuracy for each site as a function of the relative difference in streamflow between the two images shown to the annotator. Percent accuracy was computed for annotations in binned intervals of 0.1 relative flow difference. Two annotators (represented with symbols and named with 5-digit alphanumeric identifier) performed annotations across the 11 camera sites. The red line is a 4th order polynomial fit across all 11 camera sites, with equation and fit statistic shown at the bottom of the figure.

270

3.2 Modelling Results

Predictions from models trained on person-generated annotations were found to represent both individual storm events and inter-annual hydrologic changes with a satisfactory degree of fidelity, with “test-in” Kendall’s Tau values ranging from 0.60 to 0.83 (Fig. 5). This “test-in” set represents expected performance during the same period as for which annotations were created. The “test-out” period, while not available for all sites, represents predictions for a new period following an annotation as in an operational prediction with a deployed model. The value “all-in” is a mix of test/training/validation data occurring during the training period and is representative of a retrospective prediction or hindcast. The value “all” is for all

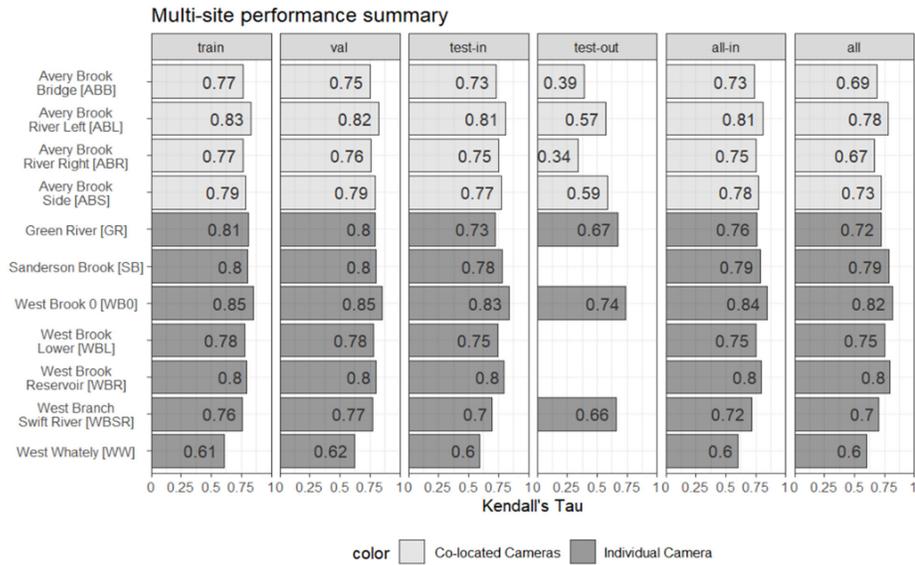
275

imagery in the dataset combined. We separately report statistics for the data splits “test-in”, “test-out”, “all-in”, and “all”. Most models have a slight decrease in performance (approximately 0.02) when comparing the training to test-in results. This decrease is a measure of overfit to the data. Green River has the greatest decrease (0.08, or 10%). A review of the annotations for this site shows a low density in annotations at the end of the training period that could account for this difference. Where available, the test-out performance is lower than test-in performance (mean decrease is 0.20), suggesting a decreased ability to generalize to new flow conditions or camera views.

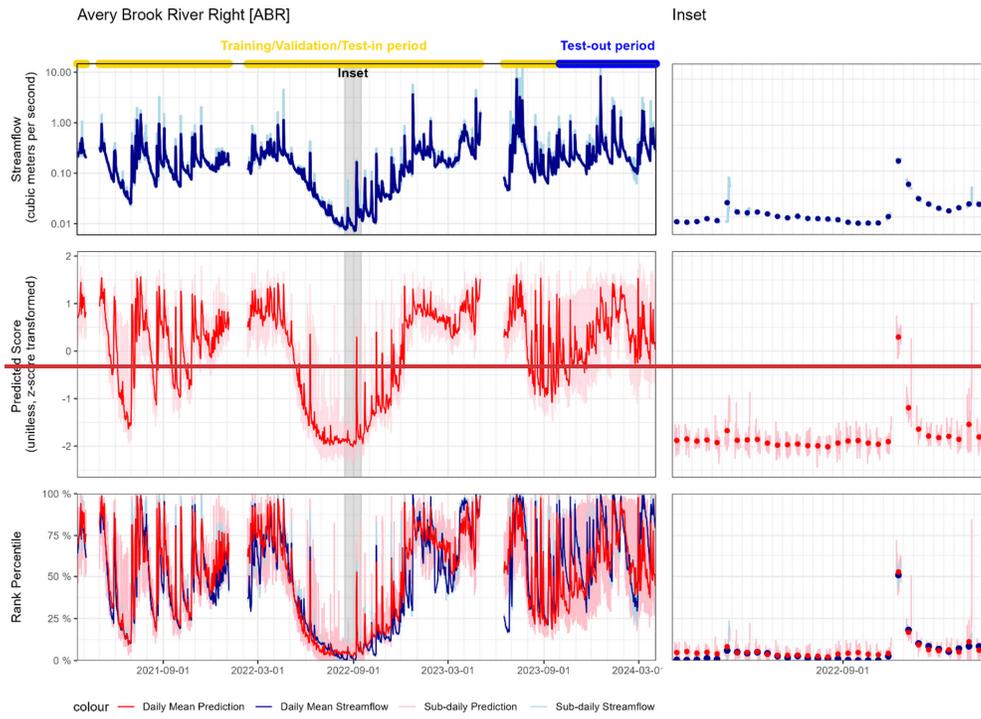
Within our camera monitoring dataset, we have several co-located cameras that were independently annotated and trained (lighter colour bars in Fig. 5). Four co-located cameras exhibited similar test-in performance, although a downstream-facing view had slightly lower performance than the other three. For the test-out period, two sites (Avery Brook River Left and Avery Brook Side) have much better performance than the other two. These sites have “high” camera stability and greater annotation accuracy than the other two sites. The streamflow has the similar (but not identical) coefficient of variation due to the differing monitoring timeframes among the cameras.

Model prediction timeseries show a clear correspondence with observed streamflow timeseries, especially when both datasets are displayed as rank percentile units (Fig. 6; supplemental materials). Major hydrologic events such as a drought that occurred in this area from June–September of 2022 and a prolonged wet period in July–August of 2023 are visible in the estimates derived solely from the imagery model. The duration and magnitude of major hydrologic events match well between observed streamflow and model predictions. Short-lived peaks from individual storm events are also well-characterized by their timing and general magnitude.

Model performance of the “test-in” set, annotation performance, flow variability, and camera stability were found to be highly interrelated (Fig. 7). Positive correlations were observed between flow variability and annotation accuracy (Panel A), flow variability and model performance (Panel C), annotation accuracy and model performance (Panel D). West Whately is an outlier to some extent; we report Pearson’s correlation coefficients and p-values with and without this camera site. The relationship between annotation accuracy and model performance (Panel D) has the highest correlation and is least affected by the outlier presence. Camera stability, a categorical variable, was weakly related to annotation accuracy (Panel B). The Kruskal-Wallis test indicates that the annotation performance is non-identical across the three stability classes at the 0.05 significance level. The *post hoc* Dunn’s pairwise multiple comparison test shows the only significant difference is between the “high” stability and “medium” stability classes. The Kruskal-Wallis test indicates there is no significant difference in Kendall’s Tau among the stability classes (Panel E). Among the four cameras located on the same stream reach (shown with lighter shading), the highest performance in annotation accuracy and prediction Kendall’s Tau was observed for Avery Brook River Left, which had a highly stable camera.



310 **Figure 5** – Summary of model performance, as defined by Kendall's Tau correlation, between observed and estimated streamflow percentile. Results are presented for 11 sites, 4 of which are co-located. Site abbreviations shown in brackets. Results are presented for six different sets of the data. The set "test-in" represents unseen images coincident with the training period. The set "test-out", which is not available at all locations, represents unseen images following the training period.



315

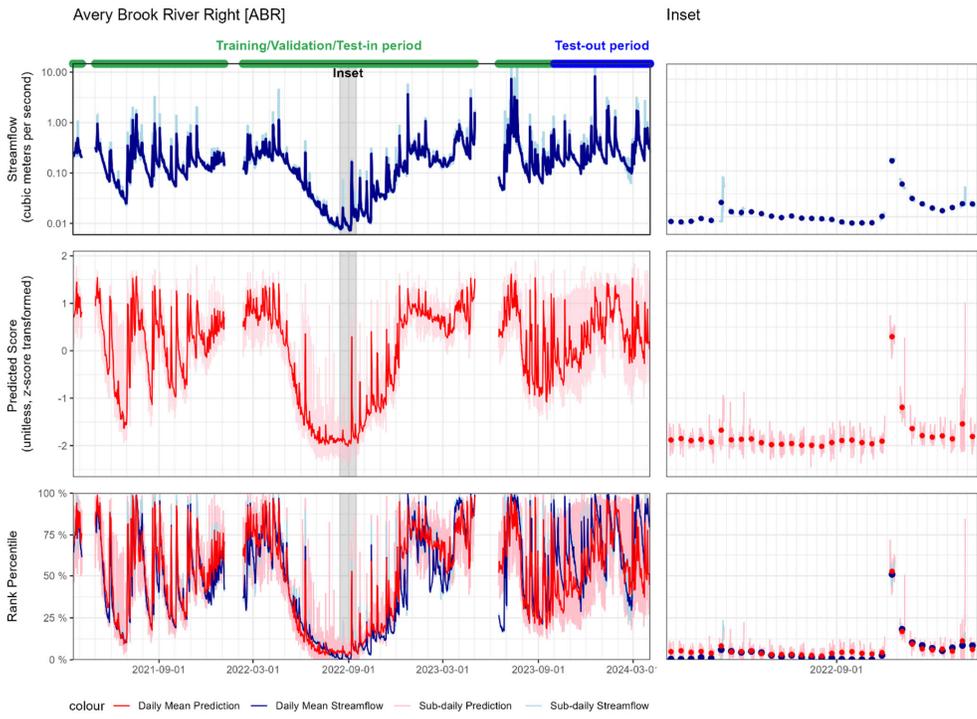
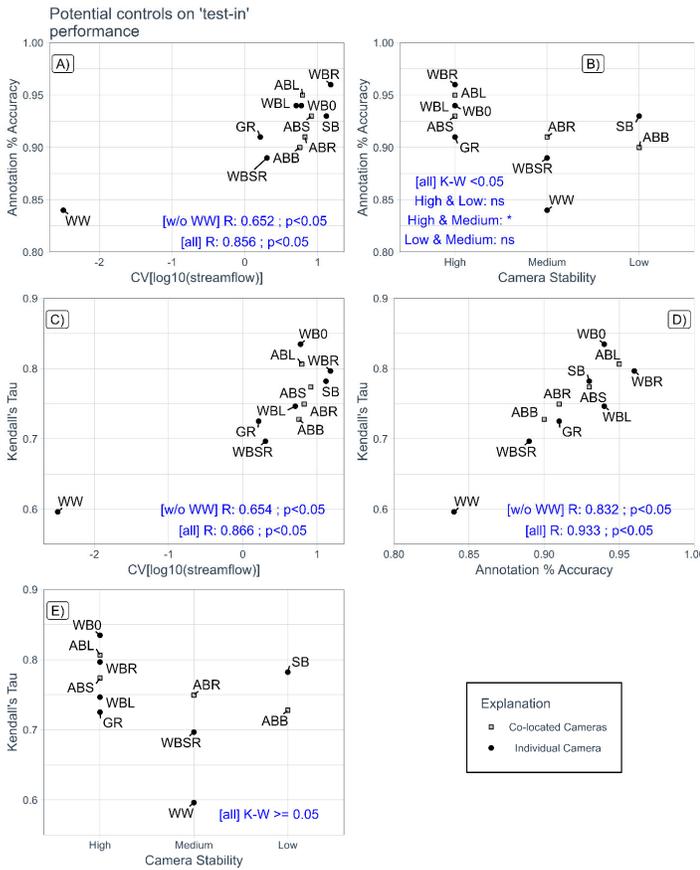


Figure 6 – Timeseries prediction at a single site representing intermediate model performance. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines. Prediction timeseries for all sites are shown in the supplemental materials.

320



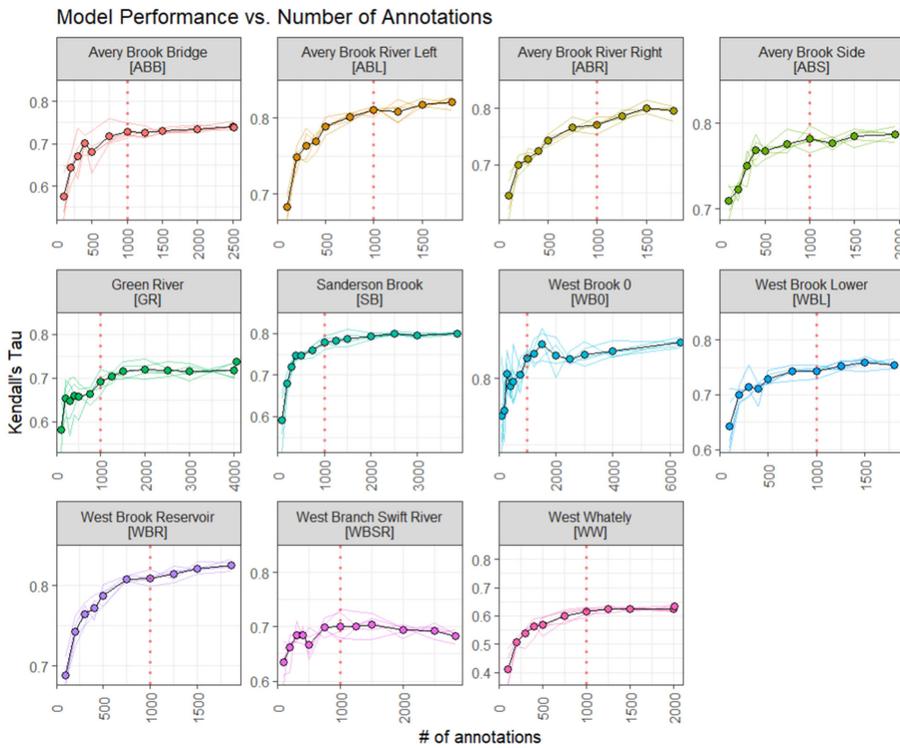
325 **Figure 7** – Relationships between flow variability and annotation accuracy (panel A), camera stability and annotation accuracy (panel B), flow variability and model performance (panel C), annotation accuracy and model performance (panel D), and model performance and camera stability category (panel E). Flow Variability is quantified with the coefficient of variation of log-transformed streamflow. Model performance is the “test-in” split. Point labels refer to site number listed in Table 1. The four co-located cameras are indicated with light grey square symbols. Panels A, C, and D have text indicating the Pearson’s correlation coefficient and significance at the $p < 0.05$ level; values are provided for without the West Whately site (“w/o WW”) and for all sites (“all”). Panels B and E have text with the Kruskal-Wallis significance test at the $p < 0.05$ level. Where significant, the post hoc Dunn’s pairwise multiple comparison test is performed. An asterisk indicates significance at the $p < 0.05$ level while “ns” indicates not significant at that level.

330

3.3 Sensitivity analysis

335 The sensitivity analysis we employed allowed us to examine the relationship between number of annotations and model performance. For most sites, “test-in” model performance improves significantly from 100 annotations to 500 annotations, as the models learn more relevant features for inferring relative streamflow (Fig. 8). Generally, the spread of individual iterations (n = 5) was small relative to the performance improvement associated with increasing annotations. The model performance plateaus around 1000 annotations for most sites. Beyond this point, additional annotations offer minimal gains, suggesting that the model is not extracting further useful information from the additional annotations.

340



345 **Figure 8** - Model performance as function of the number of annotations used to train the model. Coloured lines indicate individual scaling experiments (n=5), the black line and points indicate mean performance. Dotted vertical line shows 1000 annotations. Axis limits vary among panels. Performance computed on daytime (7am – 7pm) photos only. Subplots labelled with site name and number described in Table 1.

4 Discussion

We find that a low-cost methodology for monitoring relative streamflow dynamics in headwater streams is effective at characterizing interannual hydrologic events and short-term storm responses at the stream sites within our study. Based on our encouraging results, we anticipate the approach will provide a valuable alternative to traditional stream gaging methods when relative streamflow dynamics information is needed but the streamflow discharge is not required. The person-generated annotation, model performance patterns, and sensitivity analysis performed in this study have implications for how we refine the modelling approach and provide guidance to users as this platform evolves.

This study was our first insight into annotator accuracy. In a previous study outlining the SRE method (Gupta et al., 2022), the ability to correctly rank the image pair was varied systematically using simulations. In that study, in addition to a perfect annotator [that always ranked the image pair correctly](#), the authors simulated annotations with varying ability to discern between [“same” apparent streamflow and “left” or “right” streamflow differences in the photo pair](#). The thresholds they tested included 10%, 20%, and 50% of the lesser discharge. The authors found a less discerning annotator had to perform more annotations to train a model that reached similar performance as one trained on annotations from a more discerning annotator. However, these annotators could not make mistakes; incorrect labels were not introduced. Conversely, in this study, people performed annotations. While the overall accuracy of the annotators at individual sites ranges from 84% - 96%, these accuracy statistics obscure another feature of annotation – annotators are near perfect at distinguishing large differences in flow and less accurate at distinguishing small differences in flow. Even when provided with a “same” button and a “don’t know” button, annotators make mistakes at small differences in flow. This is likely due to the difficulty of the task in the presence of camera angle shifts, obscuring vegetation, changes in channel morphology, and the fact that it is simply difficult to discern small differences in streamflow visually. Annotation performance in our dataset followed similar patterns for two annotators and across 11 camera sites, such that all data could be reasonably fit with a single mathematical function (see [Fig. 4](#)). Future studies could use this function to simulate annotator performance more accurately than previous threshold-based simulations. This study primarily relied on annotations from two individuals with similar backgrounds and a single annotator worked on each site, resulting in a potential conflation of annotator and site variability. Future work using larger annotation datasets or designed common annotation sets could better assess the range of skill across individuals and backgrounds. [In this study we used streamflow gage observations to quantify annotator and model performance. Where observations are not available, annotator performance could be assessed using multiple annotators assessing the same image pairs. Model performance could be evaluated using post-hoc human review using a similar approach as annotation.](#)

This study’s models, trained with person-generated annotations, produced a timeseries of streamflow percentile estimates analogous to a relative hydrograph that can be used to monitor the timing, duration, and relative magnitude of hydrologic events ([Fig. 6; Supplemental Materials](#)). All performance metrics in this paper are provided for the original approximately 15-minute interval frequency of the imagery and streamflow data, though the timeseries plots of model predictions do show substantial sub-daily variability in streamflow percentiles. For example, at times in late 2023, daily

percentiles at site Avery Brook Bridge consistently range from nearly 25 % to 90% (Fig. 6). A review of individual images during times of high sub-daily percentile variability shows that outliers in model prediction can be introduced by the presence of sun glare on the camera, vegetation blocking the camera view, twilight conditions, fog/haze, and other factors that present the model with unfamiliar views (Supplemental Materials, section 3). Since we allow annotators to exclude photos that are obscured, the model is not trained on these images which leads to poorer performance. A focus of future work on identifying and excluding these images will likely reduce the variability at a sub-daily scale resulting from poor images. Even in the presence of these features, the daily mean values plotted correspond well with major hydrologic events, such as a drought in the summer of 2022 that affected the region, and individual storm events. Users of this modelled relative streamflow data could create daily mean values if they were interested in results at this scale. However, we report sub-daily model performance because the headwater streams that are a focus of this work are highly responsive to storm events and it is also important to capture these events to understand and characterize streamflow dynamics.

Where available, this study found lower model performance for the `test-out` period than the `test-in` period, though the degree of performance decrease varied among sites. Even for the four co-located sites on Avery Brook the decrease in model performance from `test-in` to `test-out` varied substantially (Fig. 5). We believe this may be due to a combination of new camera views not seen in training and the fact that the `test-out` period often included winter which can be a period of lower performance due to snow obscuring the stream. ~~Due~~[The general approach we took may be limited in its ability to describe the magnitude of out-of-distribution streamflow in the `test-out` period, but due](#) to the limited availability of sites with `test-out` periods, we are unable to draw conclusions that might hold true for other sites. Creating models from longer paired imagery and streamflow records with more extensive `test-out` periods will support future efforts to minimize performance loss for the `test-out` period, likely through improvements in the image augmentation steps of the modelling procedure.

Model performance among sites seems to be driven by the variability of the streamflow during the monitoring period. We find that annotator and model performance at sites that have very steady flow is low relative to sites experiencing wide variation in observed streamflow. To some extent this is a consequence of the Kendall's Tau as a performance metric; where a small range in the overall data causes small fluctuations in stream discharge to manifest as large fluctuations in rank percentile. However, physical characteristics matter— for this method to perform well the stream needs to have visible changes in streamflow during the training period. The site in this study with the lowest streamflow coefficient of variation, West Whately, also had a very low stream depth such that the water surface was difficult to see within a meandering channel and in the presence of leaves. Future work with more sites will be better positioned to evaluate how camera stability, flow variability, other factors affect annotation and model performance. This study refines user guidance in two important ways. First, our results suggest that sites that experience a wide range of flows (or for long enough that a wide range of flows are experienced) will have higher model performance. Second, since our simple camera stability classification has a weak association with annotator accuracy and no significant relationship with model performance, the method is robust to slight changes in camera angle and can still be used if these shifts are present. [However, the limited three-category approach in this study may limit the](#)

findings. More complex frame-tracking algorithms to quantify camera stability (i.e. Ljubičić et al., 2021) could further improve insights into the robustness of the method to camera shifts.

A key requirement of this methodology is the need for a person to perform annotations on the imagery datasets. Anecdotally, users typically annotate at an average pace of 1,000 image pairs per hour using the interface. However, in practice, annotations are typically performed in smaller batches (100-200 images per batch) with breaks in between resulting in a slower effective pace. The sensitivity analysis performed in this study helps evaluate the number of annotations to reach near-optimal model performance while not wasting annotator effort. For our available sites and annotation datasets we approximate 1,000 annotations as a reasonable guideline when creating a new model. While there is slight variability among the sites, the consistency of the shape of the curves shown in **Fig. 8** suggests that a single guideline is reasonable. The number of annotations may also be controlled by factors not included in the sensitivity analysis such as the record length and annotator accuracy. Additional sensitivity analyses, likely using synthetic annotation datasets, could further refine the guideline for how many annotations to perform when developing ranking models at new sites.

The output of the deep learning model is a relative flow percentile estimate. Although streamflow discharge (i.e. a flow rate with units of volume per time) is a more familiar metric, relative flow has value for several applications. With relative flow estimates we can 1) evaluate the duration and timing of disturbances such as drought and flood events, 2) provide inputs to statistical models such as ecological population models that may not require absolute streamflow accuracy, 3) establish or confirm relationships between streamflow at a study reach and at other nearby locations, 4) evaluate the ability of hydrologic models to simulate streamflow dynamics at a study reach, 5) provide the basis for counting the exceedance of site-relevant thresholds (for example, the number of times a roadway inundated)– or the number of times an intermittent stream is active). These outputs are aligned with the work of other authors to use semiquantitative observations to study headwater streams, for example stream connectivity (Bellucci et al., 2020; Kaplan et al., 2019). Nevertheless, some applications require absolute flow, and in future work we intend to explore approaches to transform relative flow estimates produced by the SRE model into absolute streamflow discharge estimates, either by periodically measuring discharge at the site or by using discharge data from nearby locations (if available). For now, we intend to communicate the appropriate use of these relative percentile estimates and avoid implying that streamflow discharge is produced by this work.

Because our study reports relative rather than absolute streamflow, it is difficult to directly compare our model performance against other similar work. We report our performance with the rank-based Kendall's Tau value, which is analogous to a nonparametric R^2 value appropriate to our model outputs. Similar studies using timelapse camera imagery to monitor rivers focus on reproducing point-in-time stage observations, often using in-channel calibration targets such as staff gages (Chapman et al., 2022; Eltner et al., 2018; Gilmore et al., 2013; Kim et al., 2011; Lin et al., 2018; Nguyen et al., 2009). These studies vary in approach, though typical steps include identifying the target and water surface, performing an orthorectification of the image into real world space, and conducting a measurement of a visual target. Typically, authors report sub-centimetre level accuracy. For example, a field study of uncertainty of one system reported ± 5 mm accuracy at the 90% confidence interval in a tidal marsh environment with tranquil waters, though the authors noted this system was unsuited to

fast-moving turbulent water such as the mountainous headwater streams in our dataset (Birgand et al., 2022). A deep-learning water segmentation-based approach reported Spearman correlations between independent stage measurements ranging from 0.57 to 0.94 at a single well-characterized gage site in eastern Germany (Eltner et al., 2021). We note these performance metrics reported by other similar studies, though due to differences in the model outputs our performance metrics are not directly comparable. Where evaluated in the field, most similar studies report results for single sites and/or for durations of less than 1 year (Birgand et al., 2022; Eltner et al., 2021; Leduc et al., 2018; Liu and Huang, 2024; Schoener, 2018), making this study's multi-year monitoring of 11 camera sites a comparatively robust representation of model performance.

This work, while promising, is limited in a few important ways. Primarily, this system is not (and is not intended to be) a replacement for high accuracy stream stage or discharge measurements that are required for many applications such as computing streamflow trend, calculating nutrient loads, or supporting water management decision making. Users of this system must understand the relative nature of the results and determine if relative streamflow hydrographs are suitable for their application; we envision suitable applications to include habitat characterization, aquatic species population dynamics modelling, refining process understanding in small catchment studies, intermittent stream monitoring, and characterizing event (i.e. flood or drought) timing. In this study, model training and prediction is limited to daytime imagery, which we defined simply as between 7am and 7pm local time. While these cameras also have infrared flash that illuminates the channel, the degree to which the scene is visible at night varies significantly between sites. The imagery at night becomes greyscale and we expect that different portions of the imagery become important for a model. It is unclear if nighttime imagery is best modelled with both day and nighttime imagery or if a night-only model should be trained, and future work may investigate this. We also noticed that lens fog, camera glare, vegetation blockages, and other visual impediments had a negative impact on model performance. When present, these image issues typically resulted in abrupt high or low outliers in model score. For this analysis we retained these predictions as part of the overall evaluation. We expect computer vision algorithms to detect and remove these images which would further improve model performance. Data collection on the Flow Photo Explorer platform enables users to flag "bad" images during data annotation, which will enable us to develop outlier detection algorithms for this purpose.

5 Conclusions

The camera-based methodology discussed here offers a novel approach to estimating [relative](#) streamflow. Its low cost and effort requirements should make it feasible to create dense observation networks to fill gaps in existing streamflow monitoring observations and thereby improve understanding of [relative](#) streamflow dynamics in headwater streams. While currently limited to estimates of relative streamflow trained as single-site models, we expect continued improvements that will expand the applicability and improve the ease of training models for new locations. The purpose of this paper was to answer questions regarding based on an initial set of monitoring stations. These findings will guide further development of the Flow

Photo Explorer integrated web platform that allows users to upload, annotate, model, and interpret headwater stream imagery.

480 To summarize, this study answers the following questions:

5.—How accurate are people at ranking images by streamflow?

485 ①. Overall annotation accuracy of ~~left/right selection~~ **image pair ranking** ranged from 84% to 96% (average of 92.2%) among the 11 camera sites. While limited to primarily two individuals, we see that our annotators are nearly 100% accurate at ranking stream image pairs when there are large differences in observed streamflow. Small differences in streamflow between image pairs were more difficult for the annotators to identify. Due to consistency among sites, the accuracy of person-generated streamflow annotations used in this study can be reasonably simulated with a single globally fit equation.

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

490 6.—How accurate are image-derived relative hydrographs developed using person-generated annotations?

495 ②. Kendall's Tau values for streamflow percentile predictions ranged from 0.6 to 0.83 for unannotated days within the training period. These represent the retrospective model performance. Lower performance was observed for predictions on data collected after the training period, which may have a different distribution of streamflow or changes to the image scene. Where available, Kendall's Tau values for the post-training period range from 0.34 to 0.74.

Formatted: Indent: Left: 0.5", No bullets or numbering

③

7.—Which factors influence ranking model accuracy that and indicate which unmonitored catchments would be suitable for low-cost camera monitoring?

500 ③. The primary factor describing among-site differences in performance was streamflow variability. Describing relative streamflow changes in streams with steady flow was challenging, in part due to our relative (percentile-based) metrics of performance. We expect better performance for streams that exhibit large stage variations, are seasonally dry, or have large seasonal variations in flow.

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

8.—How many person-generated annotations are required to achieve stable ranking model performance?

505 ④. An experiment indicated that for most sites there were diminishing improvements in performance after about 1,000 pairwise annotations. We therefore conclude this is a reasonable minimum number of annotations to develop a ranking model.

Formatted: Numbered + Level: 1 + Numbering Style: 1, 2, 3, ... + Start at: 1 + Alignment: Left + Aligned at: 0.25" + Indent at: 0.5"

Code Availability

Modelling code is provided at this GitHub code repository: <https://github.com/EcoSHEDS/fpe-model> (fpe-model v0.9.0).

510 **Data Availability**

The imagery, streamflow data, and model results used in this study are publicly visible on webpage (<https://www.usgs.gov/apps/ecosheds/fpe>). Streamflow data were originally reported in a U.S. Geological Survey (USGS) data release (Fair et al., 2025). Model predictions, annotation data, and sensitivity analysis data are also available as a USGS data release (Goodling et al., 2025).

515 **Author contribution**

BL, JF, JW, AG, and PG conceptualized the study; MH and TD collected the data; JW developed the web platform; JW and AG performed modelling; PG, JW, and AG performed data analysis; PG and AG wrote the manuscript draft; PG created figures; JF, BL, and AG edited the manuscript.

Competing interests

520 The authors declare that they have no conflict of interest.

Disclaimer

This work has not been formally reviewed by the U.S. Environmental Protection Agency (US EPA). The views expressed in this document do not necessarily reflect those of the US EPA. US EPA does not endorse any products or commercial services mentioned in this publication. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

525

Acknowledgements

The authors thank Josie Pilchik and Ethan Yu for their assistance with preparing the annotation datasets used in this study. We thank US Geological Survey colleagues William Farmer, Brian Pellerin, Jeffrey Baldock, and Timothy Lambert for reviews that improved the manuscript. We thank two anonymous reviewers and the journal editor (Markus Weiler) for constructive feedback that also strengthened the manuscript.

530

Financial Statement

Funding for this project is provided through the U.S. Geological Survey (USGS) Next Generation Water Observing System (NGWOS) Research and Development Program. This article was developed in part under a Professional Services Contract (68HE0B24P0246) awarded by the U.S. Environmental Protection Agency (US EPA) to Walker Environmental Research,

535 LLC with funding from the US EPA Regional-ORD Applied Research (ROAR) program. Model development work was advanced through a collaborative agreement between the Microsoft Corporation AI For Good Lab and the USGS.

References

- 540 Bellucci, C. J., Becker, M. E., Czarnowski, M., and Fitting, C.: A novel method to evaluate stream connectivity using trail cameras, *River Research & Apps*, 36, 1504–1514, <https://doi.org/10.1002/rra.3689>, 2020.
- Benda, L., Poff, N. L., Miller, D., Dunne, T., Reeves, G., Pess, G., and Pollock, M.: The Network Dynamics Hypothesis: How Channel Networks Structure Riverine Habitats, *BioScience*, 54, 413, [https://doi.org/10.1641/0006-3568\(2004\)054\[0413:TNDHHC\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0413:TNDHHC]2.0.CO;2), 2004.
- 545 Birgand, F., Chapman, K., Hazra, A., Gilmore, T., Etheridge, R., and Staicu, A.-M.: Field performance of the GaugeCam image-based water level measurement system, *PLOS Water*, 1, e0000032, <https://doi.org/10.1371/journal.pwat.0000032>, 2022.
- Briggs, M. A., Lane, J. W., Snyder, C. D., White, E. A., Johnson, Z. C., Nelms, D. L., and Hitt, N. P.: Shallow bedrock limits groundwater seepage-based headwater climate refugia, *Limnologia*, 68, 142–156, <https://doi.org/10.1016/j.limno.2017.02.005>, 2018.
- 550 Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G.: Learning to rank using gradient descent, in: *Proceedings of the 22nd International Conference on Machine Learning*, New York, NY, USA, event-place: Bonn, Germany, 89–96, <https://doi.org/10.1145/1102351.1102363>, 2005.
- Carlisle, D. M., Grantham, T. E., Eng, K., and Wolock, D. M.: Biological relevance of streamflow metrics: regional and national perspectives, *Freshwater Science*, 36, 927–940, <https://doi.org/10.1086/694913>, 2017.
- Chapman, K. W., Gilmore, T. E., Chapman, C. D., Birgand, F., Mittelstet, A. R., Harner, M. J., Mehrubeoglu, M., and Stranzl, J. E.: Technical Note: Open-Source Software for Water-Level Measurement in Images With a Calibration Target, *Water Resources Research*, 58, e2022WR033203, <https://doi.org/10.1029/2022WR033203>, 2022.
- 560 Colvin, S. A. R., Sullivan, S. M. P., Shirey, P. D., Colvin, R. W., Winemiller, K. O., Hughes, R. M., Fausch, K. D., Infante, D. M., Olden, J. D., Bestgen, K. R., Danehy, R. J., and Eby, L.: Headwater Streams and Wetlands are Critical for Sustaining Fish, Fisheries, and Ecosystem Services, *Fisheries*, 44, 73–91, <https://doi.org/10.1002/fsh.10229>, 2019.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Kai Li, and Li Fei-Fei: ImageNet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops)*, Miami, FL, 248–255, <https://doi.org/10.1109/CVPR.2009.5206848>, 2009.
- 565 Deweber, J. T., Tsang, Y., Krueger, D. M., Whittier, J. B., Wagner, T., Infante, D. M., and Whelan, G.: Importance of Understanding Landscape Biases in USGS Gage Locations: Implications and Solutions for Managers, *Fisheries*, 39, 155–163, <https://doi.org/10.1080/03632415.2014.891503>, 2014.

- Dunn, O. J.: Multiple Comparisons Using Rank Sums, *Technometrics*, 6, 241–252, <https://doi.org/10.1080/00401706.1964.10490181>, 1964.
- 570 Eltner, A., Elias, M., Sardemann, H., and Spieler, D.: Automatic Image-Based Water Stage Measurement for Long-Term Observations in Ungauged Catchments, *Water Resources Research*, 54, <https://doi.org/10.1029/2018WR023913>, 2018.
- Eltner, A., Bressan, P. O., Akiyama, T., Gonçalves, W. N., and Marcato Junior, J.: Using Deep Learning for Automatic Water Stage Measurements, *Water Resources Research*, 57, e2020WR027608, <https://doi.org/10.1029/2020WR027608>, 2021.
- 575 Fair, J. B., Bruet, C. R., Rogers, K. M., Dubreuil, T. L., Hayden, M. J., Hitt, N. P., Letcher, B. H., and Snyder, C. D.: USGS EcoDrought Stream Discharge, Gage Height and Water Temperature Data in Massachusetts (ver. 2.0, February 2025), <https://doi.org/10.5066/P9ES4RQS>, 2025.
- Gilmore, T. E., Birgand, F., and Chapman, K. W.: Source and magnitude of error in an inexpensive image-based water level measurement system, *Journal of Hydrology*, 496, 178–186, <https://doi.org/10.1016/j.jhydrol.2013.05.011>, 2013.
- 580 Goodling, P. J., Fair, J. B., Gupta, A., Walker, J., Dubreuil, T. L., Hayden, M. J., and Letcher, B.: Model Predictions, Observations, and Annotation Data for Deep Learning Models Developed To Estimate Relative Flow at 11 Massachusetts Streamflow Sites, 2017-2024, <https://doi.org/10.5066/P14LU6CQ>, 2025.
- Gupta, A., Chang, T., Walker, J., and Letcher, B.: Towards Continuous Streamflow Monitoring with Time-Lapse Cameras and Deep Learning, in: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS), COMPASS '22: ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies, Seattle WA USA, 353–363, <https://doi.org/10.1145/3530190.3534805>, 2022.
- 585 He, K., Zhang, X., Ren, S., and Sun, J.: Deep Residual Learning for Image Recognition, <https://doi.org/10.48550/ARXIV.1512.03385>, 2015.
- Hitt, N. P., Landsman, A. P., and Raesly, R. L.: Life history strategies of stream fishes linked to predictors of hydrologic stability, *Ecology and Evolution*, 12, e8861, <https://doi.org/10.1002/ece3.8861>, 2022.
- Horner, I., Renard, B., Le Coz, J., Branger, F., McMillan, H. K., and Pierrefeu, G.: Impact of Stage Measurement Errors on Streamflow Uncertainty, *Water Resour. Res.*, 54, 1952–1976, <https://doi.org/10.1002/2017WR022039>, 2018.
- Jaeger, K. L., Hafen, K. C., Dunham, J. B., Fritz, K. M., Kampf, S. K., Barnhart, T. B., Kaiser, K. E., Sando, R., Johnson, S. L., McShane, R. R., and Dunn, S. B.: Beyond Streamflow: Call for a National Data Repository of Streamflow Presence for Streams and Rivers in the United States, *Water*, 13, 1627, <https://doi.org/10.3390/w13121627>, 2021.
- 595 Kaplan, N. H., Sohr, E., Blume, T., and Weiler, M.: Monitoring ephemeral, intermittent and perennial streamflow: a dataset from 182 sites in the Atert catchment, Luxembourg, *Earth Syst. Sci. Data*, 11, 1363–1374, <https://doi.org/10.5194/essd-11-1363-2019>, 2019.
- Kassambara, A.: *rstatix: Pipe-Friendly Framework for Basic Statistical Tests*, 2023.
- 600 Kendall, M. G.: A NEW MEASURE OF RANK CORRELATION, *Biometrika*, 30, 81–93, <https://doi.org/10.1093/biomet/30.1-2.81>, 1938.
- Keys, T. A., Jones, C. N., Scott, D. T., and Chuquin, D.: A cost-effective image processing approach for analyzing the ecohydrology of river corridors: Image processing of fluvial ecohydrology, *Limnol. Oceanogr. Methods*, 14, 359–369, <https://doi.org/10.1002/lom3.10095>, 2016.

- 605 Kim, J., Han, Y., and Hahn, H.: Embedded implementation of image-based water-level measurement system, *IET Comput. Vis.*, 5, 125, <https://doi.org/10.1049/iet-cvi.2009.0144>, 2011.
- King, T., Hundt, S., Simonson, A., and Blasch, K.: Evaluation of Select Velocity Measurement Techniques for Estimating Discharge in Small Streams across the United States, *J American Water Resour Assoc*, 58, 1510–1530, <https://doi.org/10.1111/1752-1688.13053>, 2022.
- 610 Krabbenhoft, C. A., Allen, G. H., Lin, P., Godsey, S. E., Allen, D. C., Burrows, R. M., DelVecchia, A. G., Fritz, K. M., Shanafield, M., Burgin, A. J., Zimmer, M. A., Datry, T., Dodds, W. K., Jones, C. N., Mims, M. C., Franklin, C., Hammond, J. C., Zipper, S., Ward, A. S., Costigan, K. H., Beck, H. E., and Olden, J. D.: Assessing placement bias of the global river gauge network, *Nat Sustain*, 5, 586–592, <https://doi.org/10.1038/s41893-022-00873-0>, 2022.
- Kruskal, W. H. and Wallis, W. A.: Use of Ranks in One-Criterion Variance Analysis, *Journal of the American Statistical Association*, 47, 583–621, <https://doi.org/10.1080/01621459.1952.10483441>, 1952.
- Leduc, P., Ashmore, P., and Sjogren, D.: Technical note: Stage and water width measurement of a mountain stream using a simple time-lapse camera, *Hydrol. Earth Syst. Sci.*, 22, 1–11, <https://doi.org/10.5194/hess-22-1-2018>, 2018.
- Levin, S. B., Briggs, M. A., Foks, S. S., Goodling, P. J., Raffensperger, J. P., Rosenberry, D. O., Scholl, M. A., Tiedeman, C. R., and Webb, R. M.: Uncertainties in measuring and estimating water-budget components: Current state of the science, *WIREs Water*, 10, e1646, <https://doi.org/10.1002/wat2.1646>, 2023.
- 620 Lin, Y.-T., Lin, Y.-C., and Han, J.-Y.: Automatic water-level detection using single-camera images with varied poses, *Measurement*, 127, 167–174, <https://doi.org/10.1016/j.measurement.2018.05.100>, 2018.
- Liu, W.-C. and Huang, W.-C.: Evaluation of deep learning computer vision for water level measurements in rivers, *Heliyon*, 10, e25989, <https://doi.org/10.1016/j.heliyon.2024.e25989>, 2024.
- 625 Ljubičić, R., Strelnikova, D., Perks, M. T., Eltner, A., Peña-Haro, S., Pizarro, A., Dal Sasso, S. F., Scherling, U., Vuono, P., and Manfreda, S.: A comparison of tools and techniques for stabilising unmanned aerial system (UAS) imagery for surface flow observations, *Hydrology and Earth System Sciences*, 25, 5105–5132, <https://doi.org/10.5194/hess-25-5105-2021>, 2021.
- McKay, L., Bondelid, T., Dewald, T., Johnston, J., Moore, R., and Rea, A.: NHDPlus version 2: User guide, United States Environmental Protection Agency. https://www.epa.gov/system/files/documents/2023-04/NHDPlusV2_User_Guide.pdf. Last Accessed: May 2024, 2012.
- 630 McManamay, R. A. and DeRolph, C. R.: A stream classification system for the conterminous United States, *Sci Data*, 6, 190017, <https://doi.org/10.1038/sdata.2019.17>, 2019.
- McMillan, H., Krueger, T., and Freer, J.: Benchmarking observational uncertainties for hydrology: rainfall, river discharge and water quality: BENCHMARKING OBSERVATIONAL UNCERTAINTIES FOR HYDROLOGY, *Hydrol. Process.*, 26, 4078–4111, <https://doi.org/10.1002/hyp.9384>, 2012.
- 635 Messenger, M. L., Lehner, B., Cockburn, C., Lamouroux, N., Pella, H., Snelder, T., Tockner, K., Trautmann, T., Watt, C., and Datry, T.: Global prevalence of non-perennial rivers and streams, *Nature*, 594, 391–397, <https://doi.org/10.1038/s41586-021-03565-5>, 2021.
- 640 Moore, R. B., McKay, L. D., Rea, A. H., Bondelid, T. R., Price, C. V., Dewald, T. G., and Johnston, C. M.: User's guide for the national hydrography dataset plus (NHDPlus) high resolution, Reston, VA, <https://doi.org/10.3133/ofr20191096>, 2019.

Nguyen, L. S., Schaeli, B., Sage, D., Kayal, S., Jeanbourquin, D., Barry, D. A., and Rossi, L.: Vision-based system for the control and measurement of wastewater flow rate in sewer systems, *Water Science and Technology*, 60, 2281–2289, <https://doi.org/10.2166/wst.2009.659>, 2009.

645 Noto, S., Tauro, F., Petroselli, A., Apollonio, C., Botter, G., and Grimaldi, S.: Low-cost stage-camera system for continuous water-level monitoring in ephemeral streams, *Hydrological Sciences Journal*, 67, 1439–1448, <https://doi.org/10.1080/02626667.2022.2079415>, 2022.

R Core Team: R: A language and environment for statistical computing, 2021.

Schoener, G.: Time-Lapse Photography: Low-Cost, Low-Tech Alternative for Monitoring Flow Depth, *J. Hydrol. Eng.*, 23, 06017007, [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001616](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001616), 2018.

650 Seybold, E. C., Bergstrom, A., Jones, C. N., Burgin, A. J., Zipper, S., Godsey, S. E., Dodds, W. K., Zimmer, M. A., Shanafield, M., Datry, T., Mazor, R. D., Messenger, M. L., Olden, J. D., Ward, A., Yu, S., Kaiser, K. E., Shogren, A., and Walker, R. H.: How low can you go? Widespread challenges in measuring low stream discharge and a path forward, *Limnol Oceanogr Letters*, 8, 804–811, <https://doi.org/10.1002/lol2.10356>, 2023.

655 Shorten, C. and Khoshgoftaar, T. M.: A survey on Image Data Augmentation for Deep Learning, *J Big Data*, 6, 60, <https://doi.org/10.1186/s40537-019-0197-0>, 2019.

Steward, A. L., Von Schiller, D., Tockner, K., Marshall, J. C., and Bunn, S. E.: When the river runs dry: human and ecological values of dry riverbeds, *Frontiers in Ecol & Environ*, 10, 202–209, <https://doi.org/10.1890/110136>, 2012.

Turnipseed, D. P. and Sauer, V. B.: Discharge measurements at gaging stations, Reston, VA, <https://doi.org/10.3133/tm3A8>, 2010.

660 U.S. Geological Survey: National Water Information System, 2024. <https://doi.org/10.5066/F7P55KJN>

Vandaele, R., Dance, S. L., and Ojha, V.: Deep learning for automated river-level monitoring through river-camera images: an approach based on water segmentation and transfer learning, *Hydrol. Earth Syst. Sci.*, 25, 4435–4453, <https://doi.org/10.5194/hess-25-4435-2021>, 2021.

fpe-model v0.9.0: <https://github.com/EcoSHEDS/fpe-model>.

665

~~This information product has been peer reviewed and approved for publication as a preprint by the U.S. Geological Survey.~~

5 S1 Model Development Description

Several pre-processing steps are applied to the imagery during model training and batch inference to better allow the model to learn the relationship between image scene and apparent streamflow (**Figure S1**). Image pre-processing occurs on all images prior to prediction while image augmentations occur only on the images comprising the 80% of the training/annotation datapairs used to develop the model. The purpose of image augmentation is to allow the model to better generalize to unseen imagery. Additional detail on the pre-processing and jittering steps are provided below:

Image Pre-Processing

Resizing: Resizes the input image to a maximum height of 480 pixels while maintaining the aspect ratio. This reduces the computational complexity and memory usage of the model by lowering the number of parameters it needs to learn.

Normalization: Standardizes the pixel values by adjusting them to have a mean of zero and a standard deviation of one. This ensures model inputs are on a similar scale, which stabilizes gradient calculations during training and accelerates model convergence.

Centre-cropping: Crops the central region of the image, excluding 10% of pixels from the edges (top, bottom, left, and right). This removes extraneous information such as camera metadata imprinted on the edges of the images, ensuring the model focuses on the stream features in the central portion.

Image Augmentations

Random Horizontal Flip: Randomly flips the image horizontally with a 50% probability.

Random Rotation: Rotates the image randomly between -10° and $+10^\circ$. This accounts for minor rotational shifts that can occur due to camera slip over long deployment periods.

Random Cropping: Crops a random part of the image covering 80% of the image. This replaces centre-cropping during training and encourages the model to learn visual features from various parts of the image and makes it more robust to slight viewpoint changes that could be induced by camera movement or partial obstructions.

Colour Jittering: Randomly changes the brightness, contrast, saturation, and hue of an image to help the model become more robust to colour variations and lighting changes in real-world scenarios.

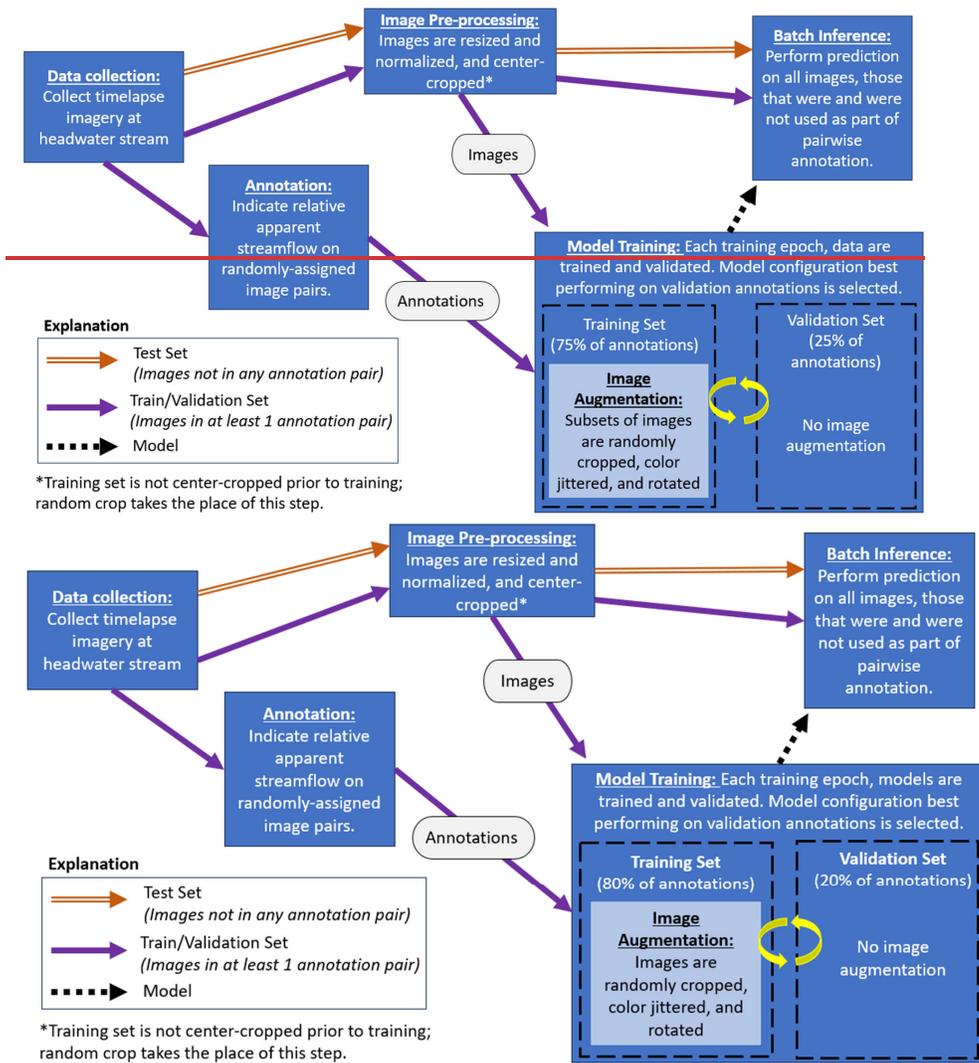


Figure S1: Diagram showing the handling of training, testing, and validation sets during model training.

Model training used the same configuration as described in (Gupta et al., 2022), with the exception of the number of epochs considered during training (raised from 15 to 20). As in that study, we used a batch size of 64, a learning rate of 0.001, a stochastic gradient descent optimizer, and a learning rate scheduler that reduces the learning rate when the validation set loss plateaus. Model training occurred on the pretrained ResNet-18 model, with the body weights frozen for the first 2 epochs and then unfrozen to allow for fine tuning for the remaining epochs. Within the 20 training epochs considered, we selected the model with the lowest validation loss as the final model. Table S1 shows the optimal number of epochs selected for each site, as well as the number of annotation pairs used for training, validation, and the proportion of annotator selections used for each site. Annotator selections of “Don’t Know” were not stored or used during model development.

Formatted: Font: Bold

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Formatted: Font: (Default) +Body (Times New Roman)

Formatted: Centered

Location ID	Station Name	Number of Training Pairs	Number of Validation Pairs	Optimal Epochs Selected During Training	% of Annotator Selections		
					LEFT	RIGHT	SAME
ABB	Avery Brook Bridge	2512	635	18	43.0	42.0	15.0
ABL	Avery Brook River Left	1817	460	18	46.1	44.5	9.4
ABR	Avery Brook River Right	1773	441	16	45.1	46.7	8.2
ABS	Avery Brook Side	1955	486	20	40.5	44.7	14.8
GR	Green River	4059	995	20	48.0	45.3	6.7
SB	Sanderson Brook	3856	965	20	45.5	44.9	9.6
WBSR	West Branch Swift River	2838	715	20	46.2	41.2	12.6
WB0	West Brook 0	6365	1588	19	46.1	45.3	8.6
WBL	West Brook Lower	1809	447	18	43.4	49.0	7.6
WBR	West Brook Reservoir	1862	463	20	40.3	45.6	14.0
WW	West Whately	2007	503	19	41.2	45.7	13.1

50 ***Table S1: Annotation pairs used in model development and number of training epochs for each model developed in this study.***

▲

Formatted: Font: Not Italic

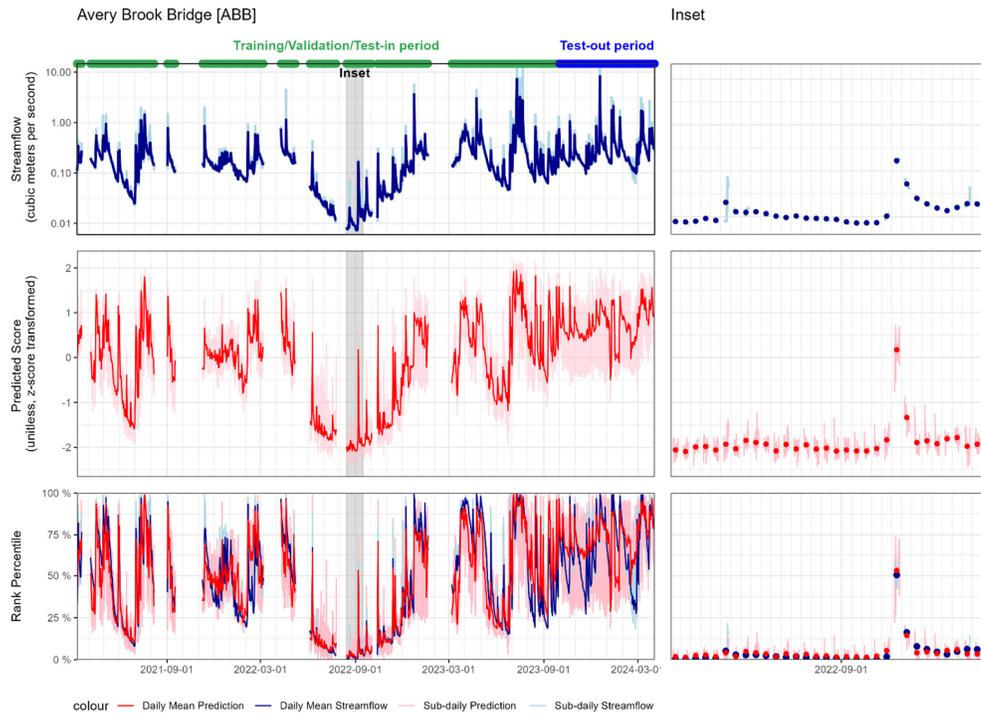
55

60

65 **S2 Timeseries Predictions for all Sites**

These figures display the timeseries of prediction for 10 camera sites and are the same as **Figure 6** in the main text, which shows the eleventh camera site. The top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

70



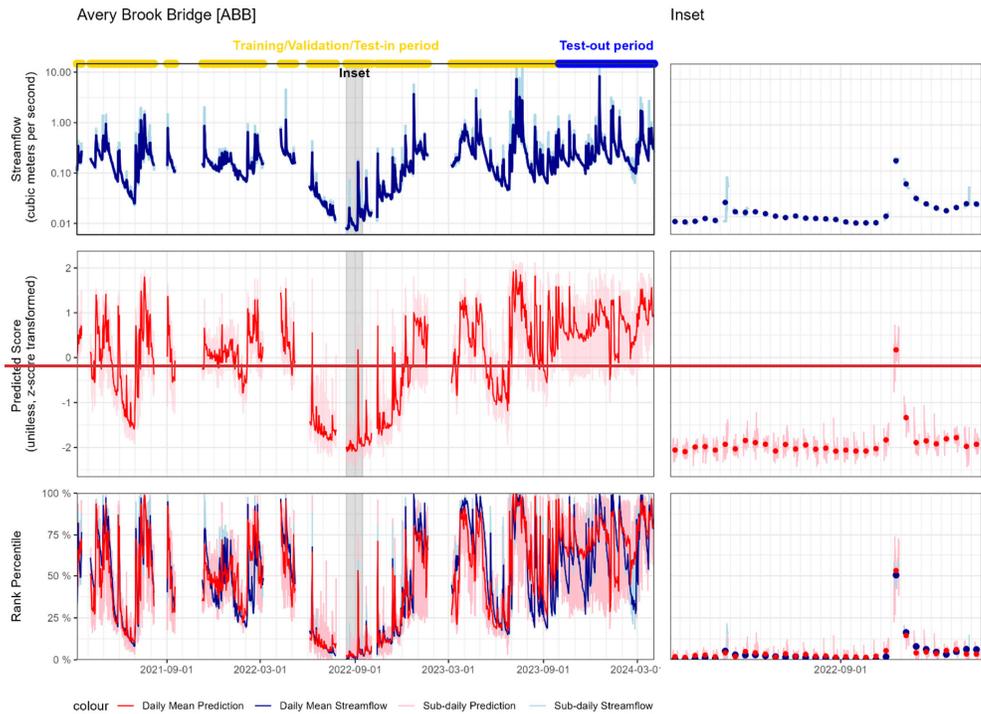
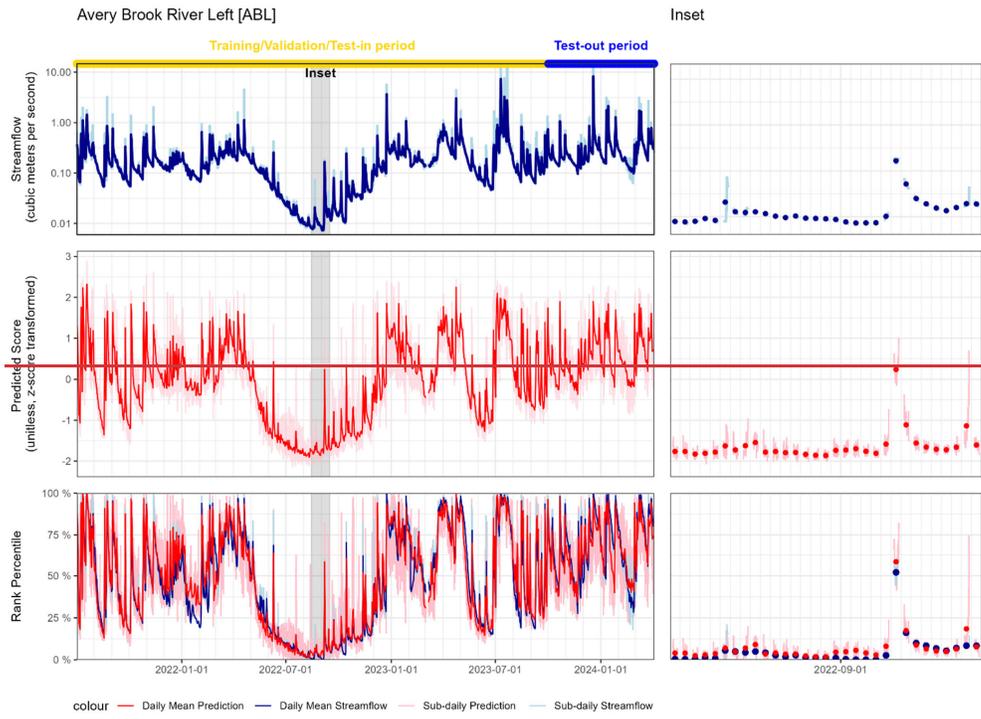


Figure S2: Timeseries prediction at the Avery Brook Bridge site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.



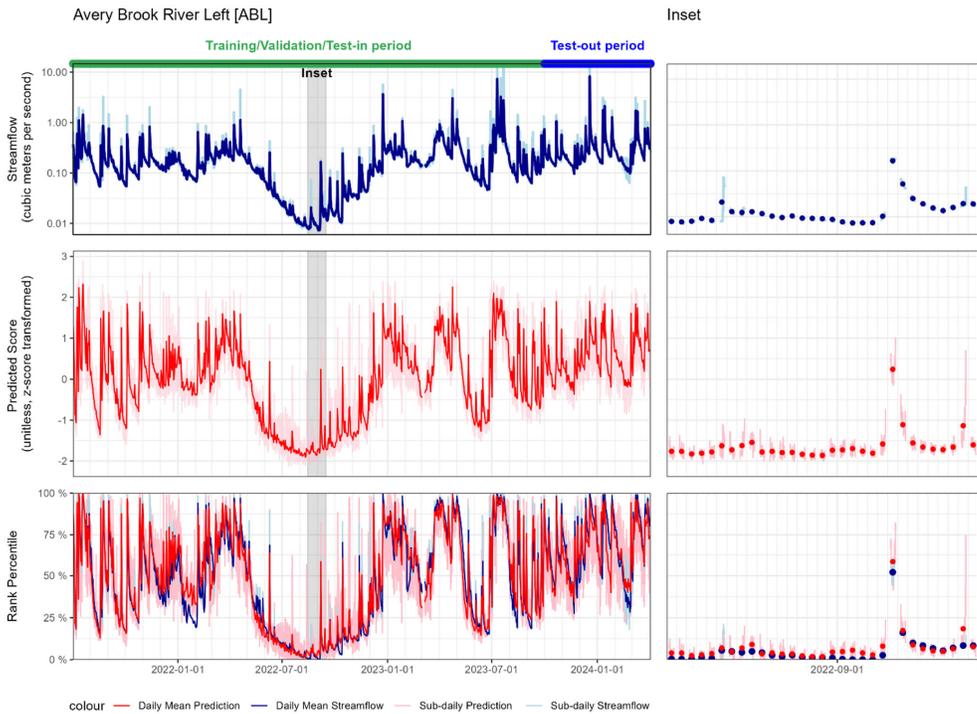
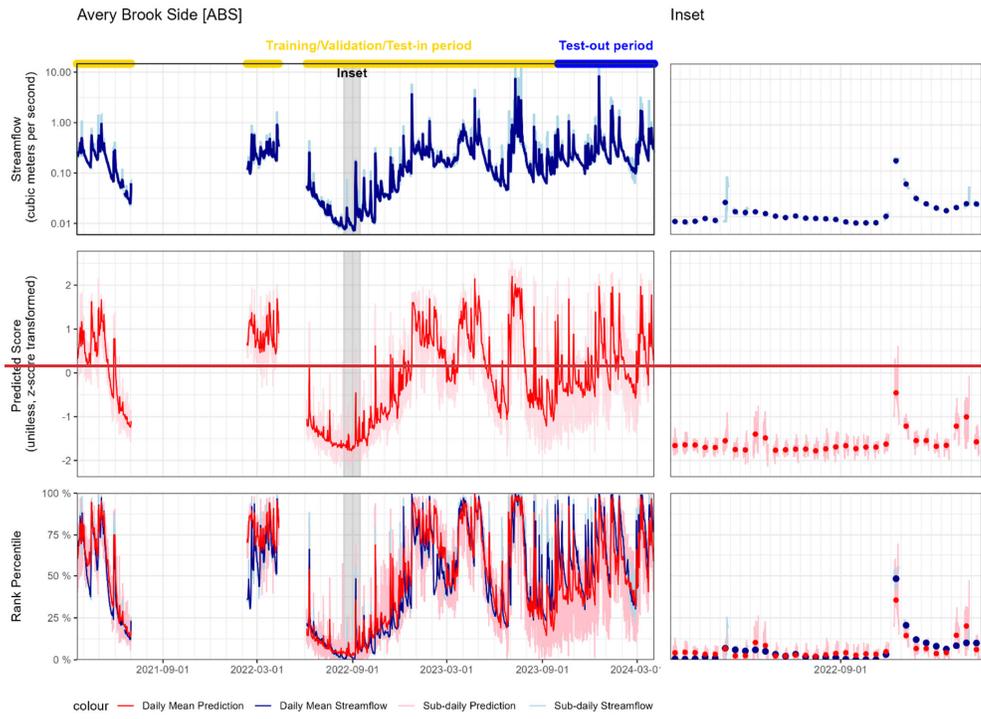


Figure S3: Timeseries prediction at the Avery Brook Left site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

85



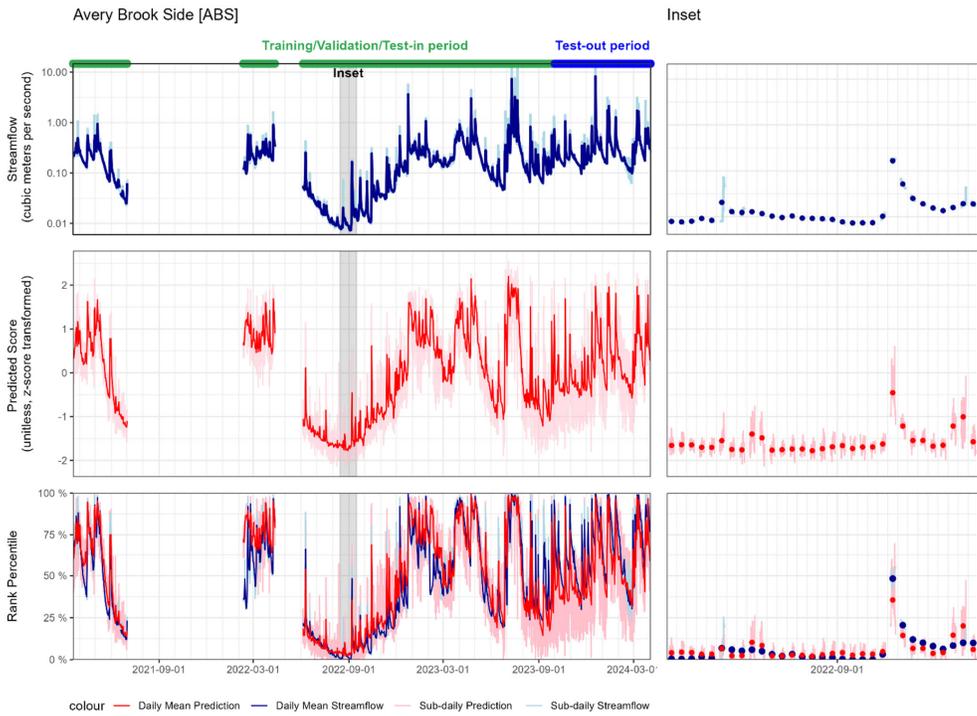
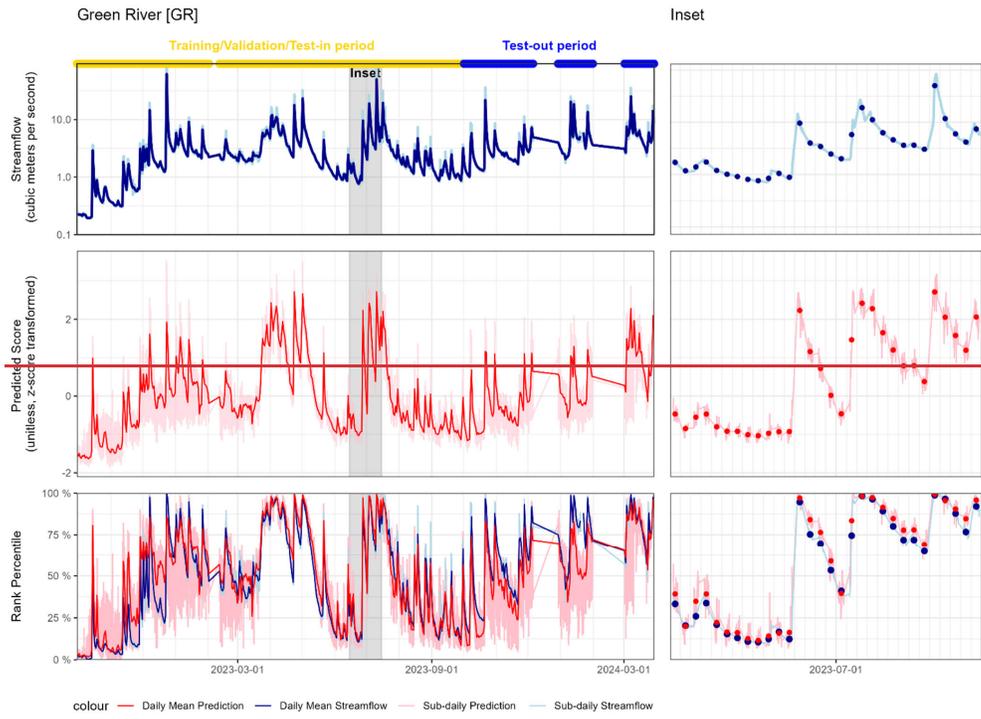


Figure S4: Timeseries prediction at the Avery Brook Side site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

90



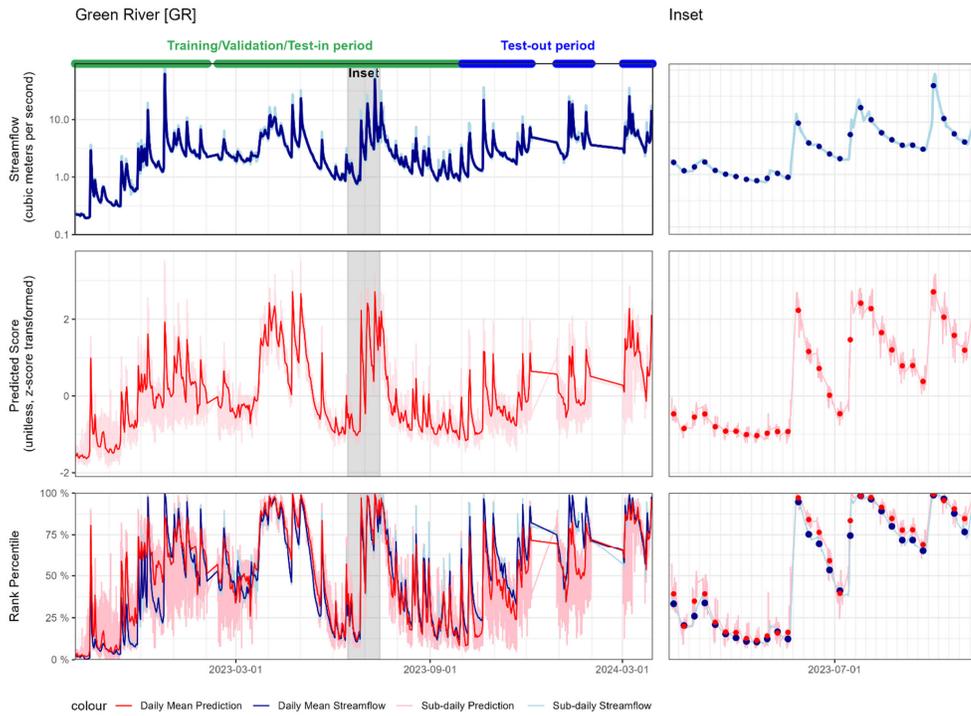
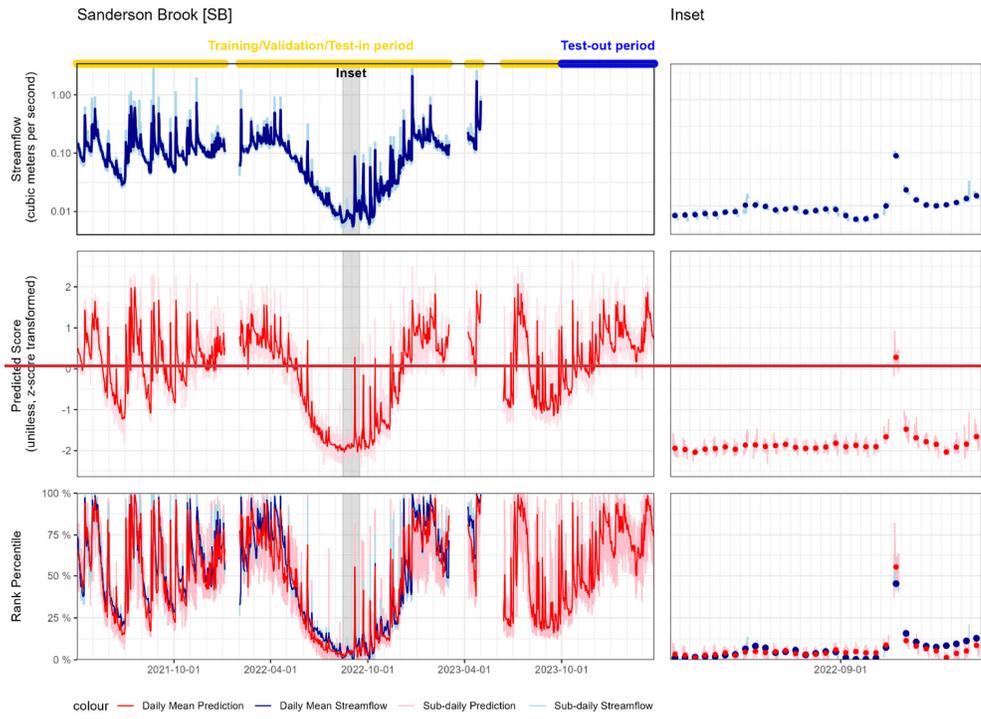
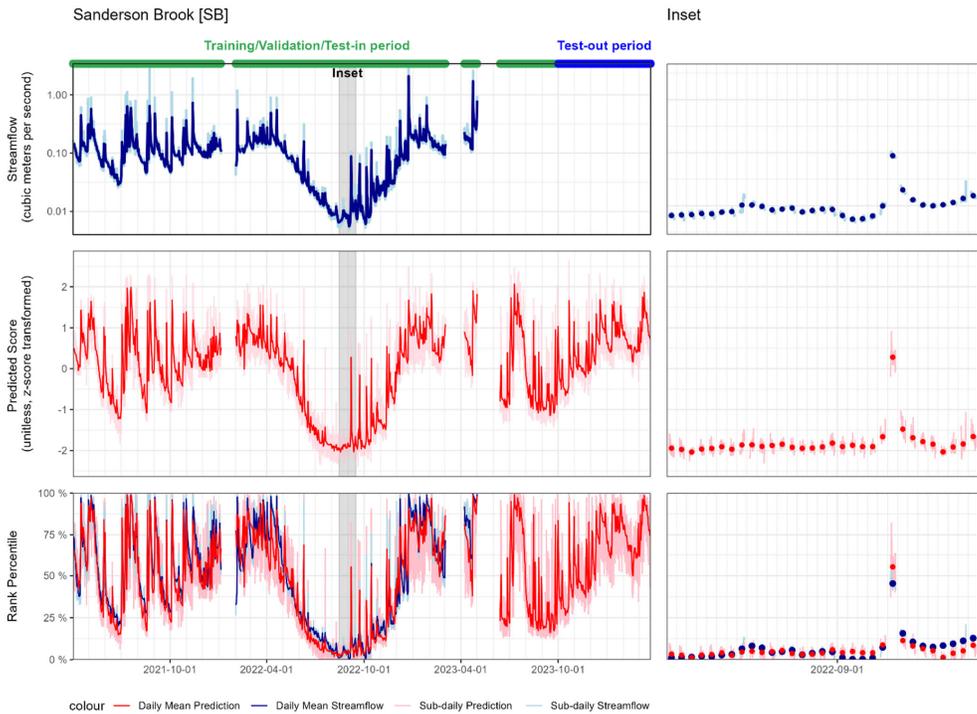


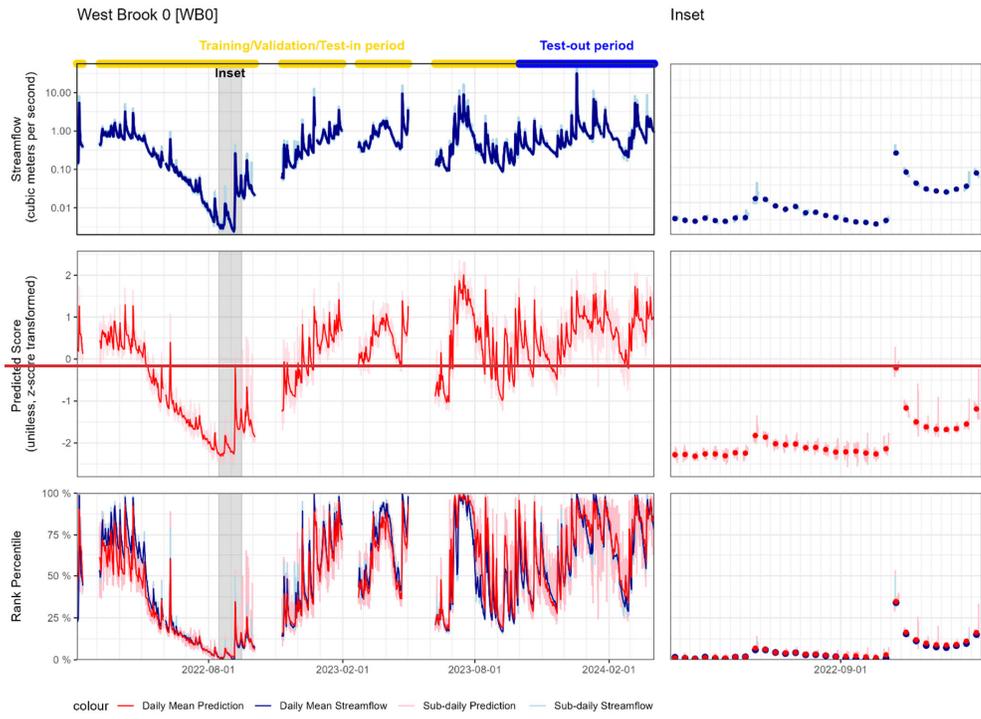
Figure S5: Timeseries prediction at the Green River site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

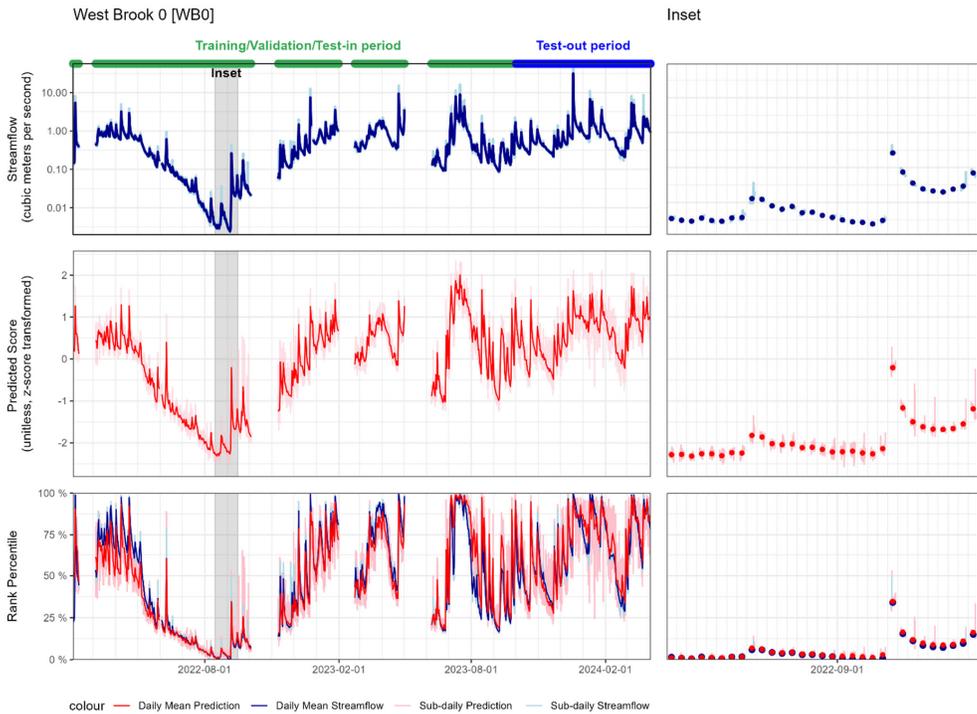
100





105 **Figure S6:** Timeseries prediction at the Sanderson Brook site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

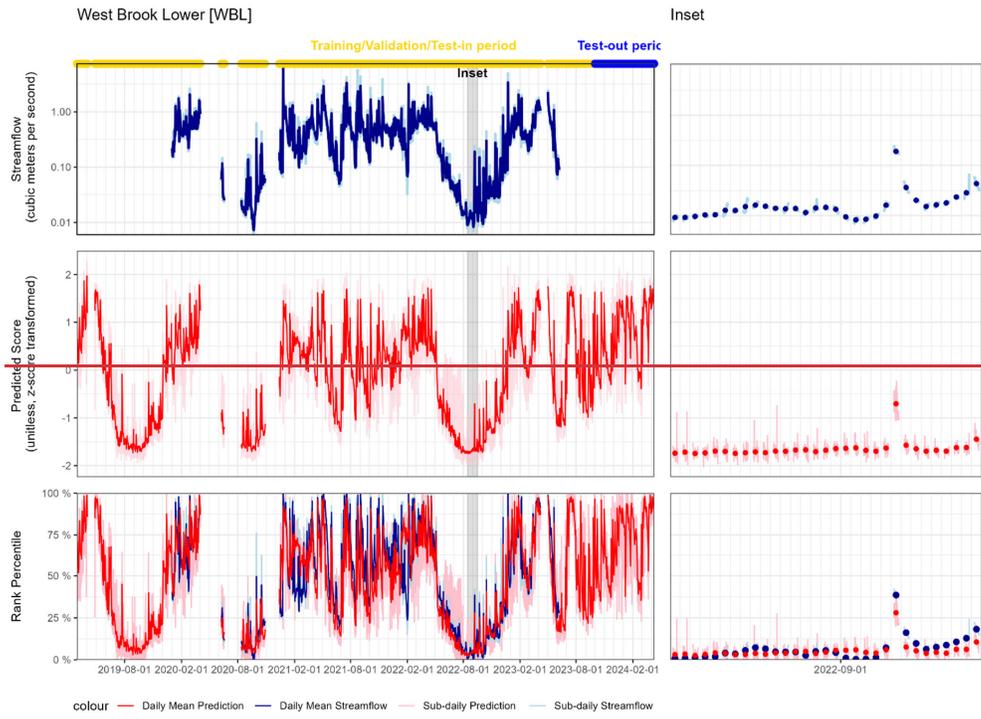




110

Figure S7: Timeseries prediction at the West Brook 0 site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

115



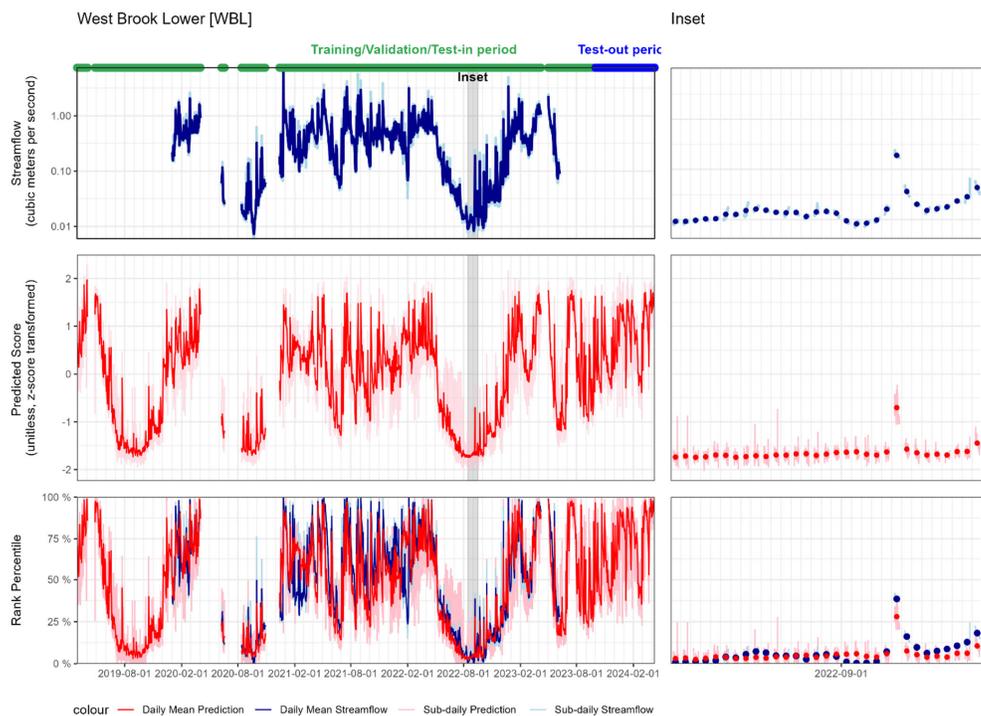
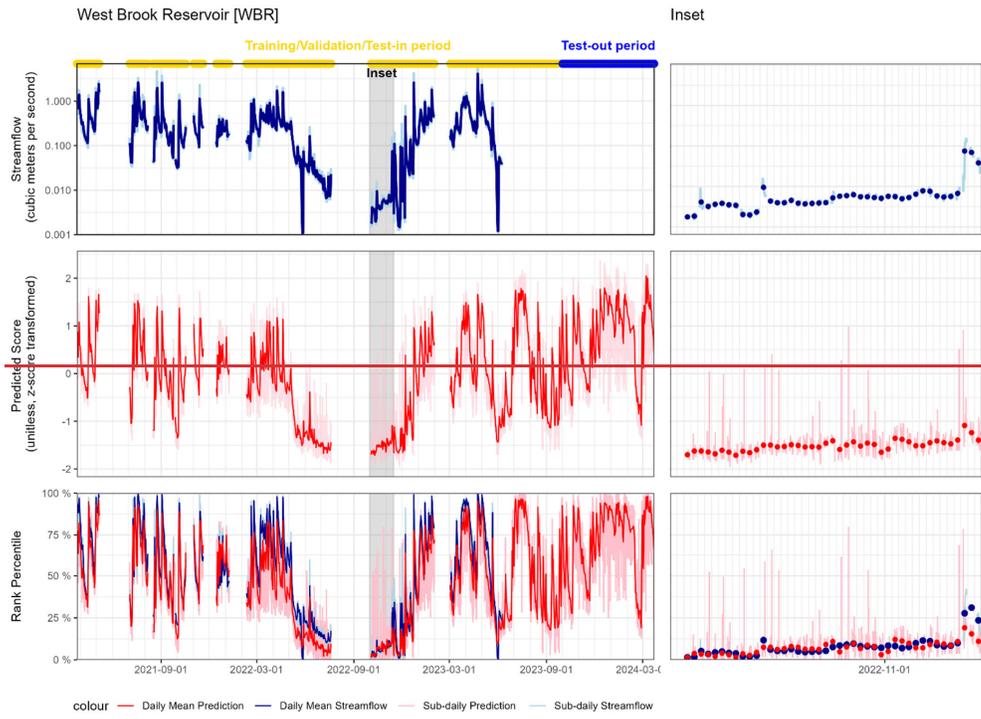
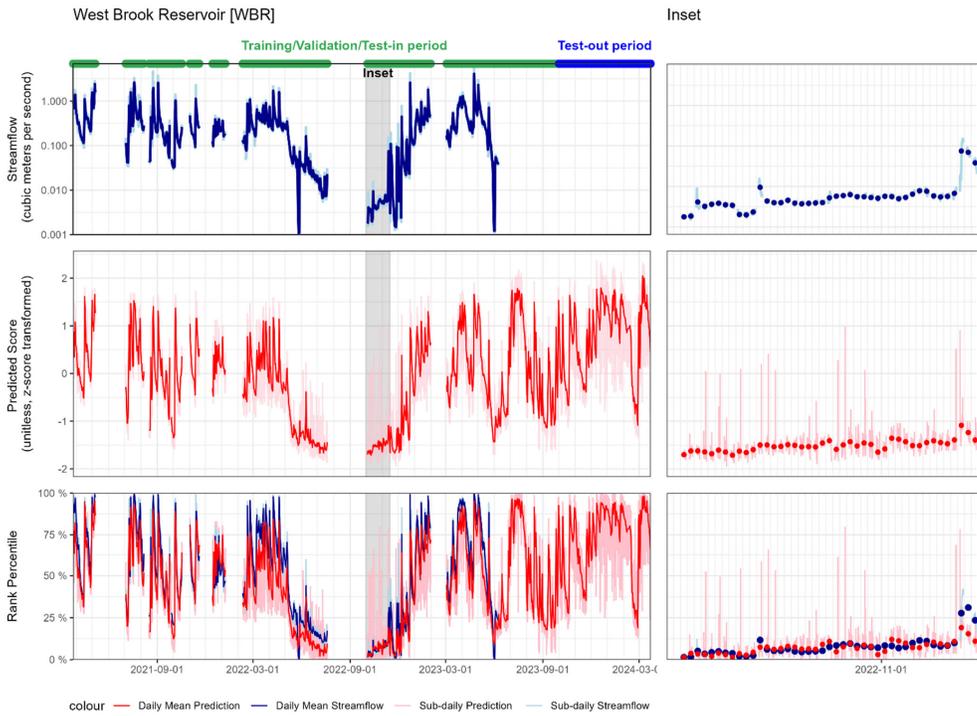


Figure S8: Timeseries prediction at the West Brook Lower site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

120

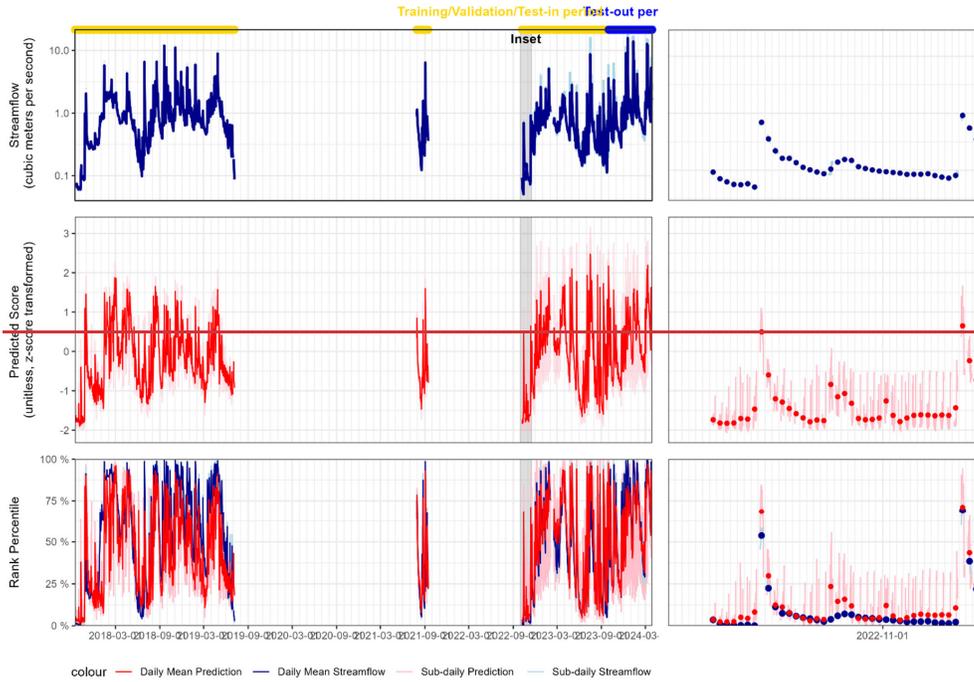




125 **Figure S9:** Timeseries prediction at the West Brook Reservoir site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

West Branch Swift River [WBSR]

Inset



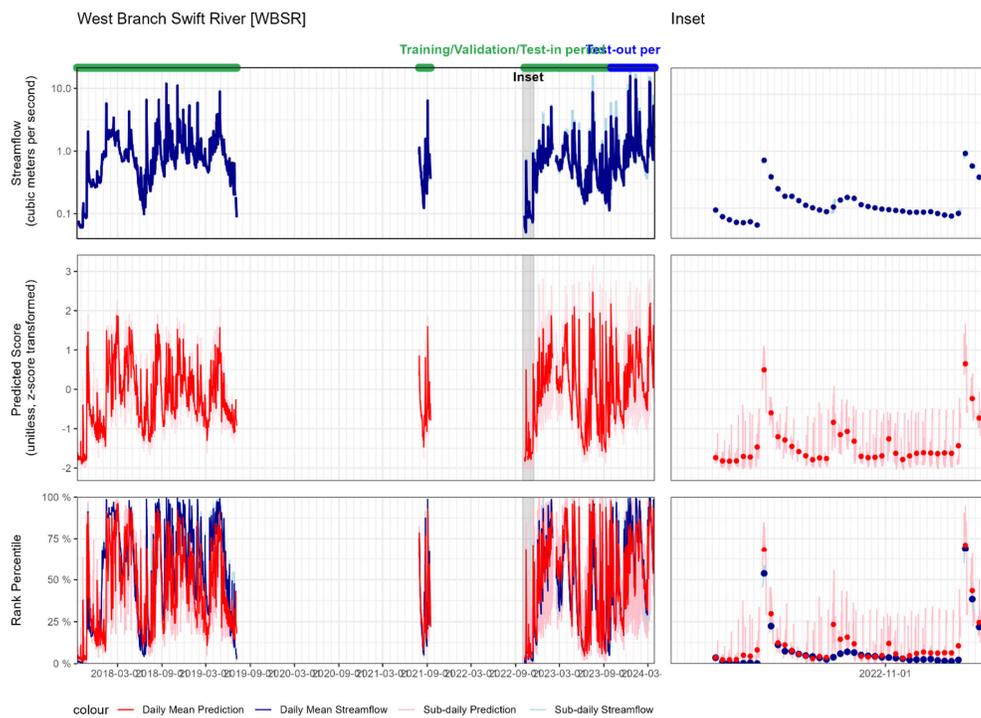
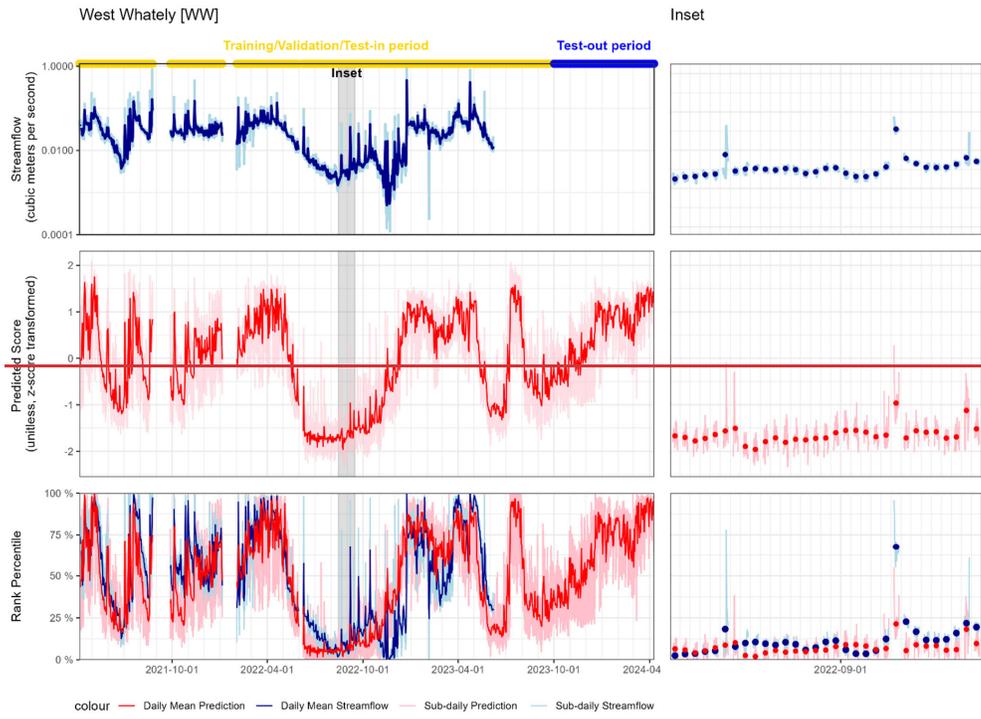


Figure S10: Timeseries prediction at the West Branch Swift River site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

135



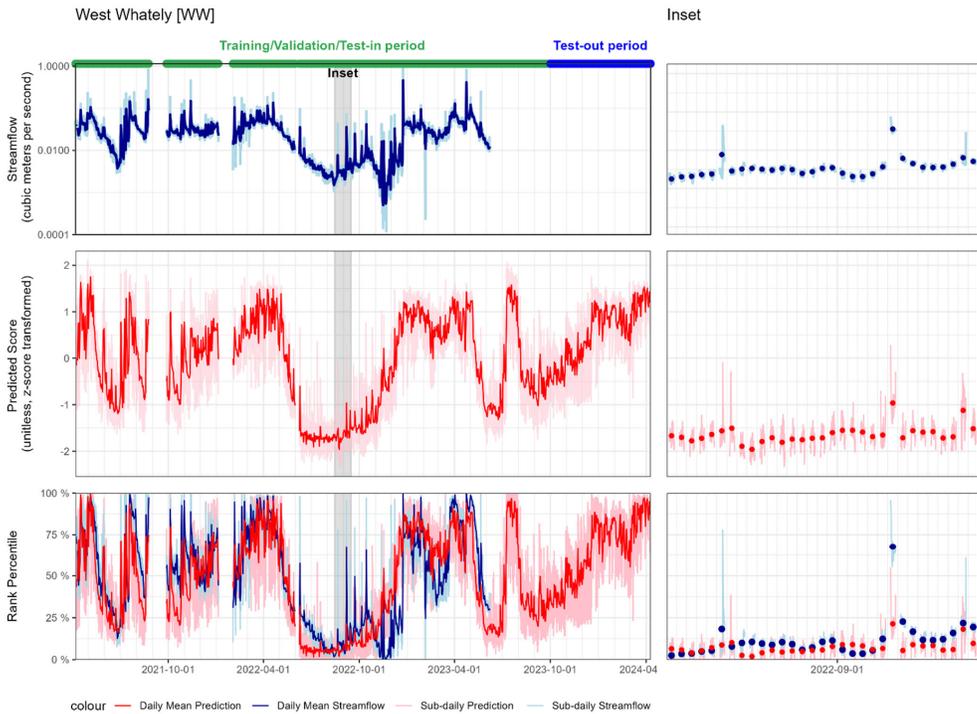


Figure S11: Timeseries prediction at the West Whately site. Top two panels show the streamflow, middle two panels show the predicted model score, bottom two panels show both when transformed to rank percentile. The left column indicates the full period of record, the right column is an inset. In the inset plots, daily means are plotted as dots and the 15-minute interval predictions are plotted with lines.

140

145

S3 Predicted vs. Observed plots for all sites (rank percentiles)

150 These figures show predicted vs. observed rank percentile for each image in the train, validation, test-in, and test-out sets for all 11 camera sites. Sites are labelled with their abbreviations which are provided in **Table 1** of the main text.

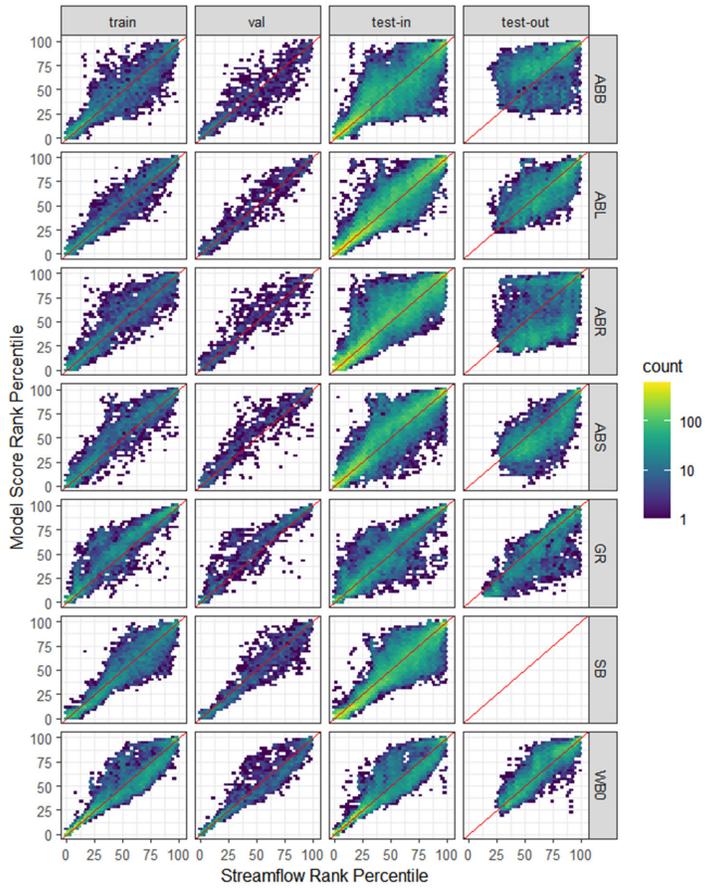


Figure S12: Predicted versus observed rank percentiles for each location and data split, shown as two-dimensional density plots. Colour indicates the number of 15-minute observations in each bin and is on a logarithmic scale. The red line is a 1:1 line. This first panel shows seven of the eleven camera sites.

155

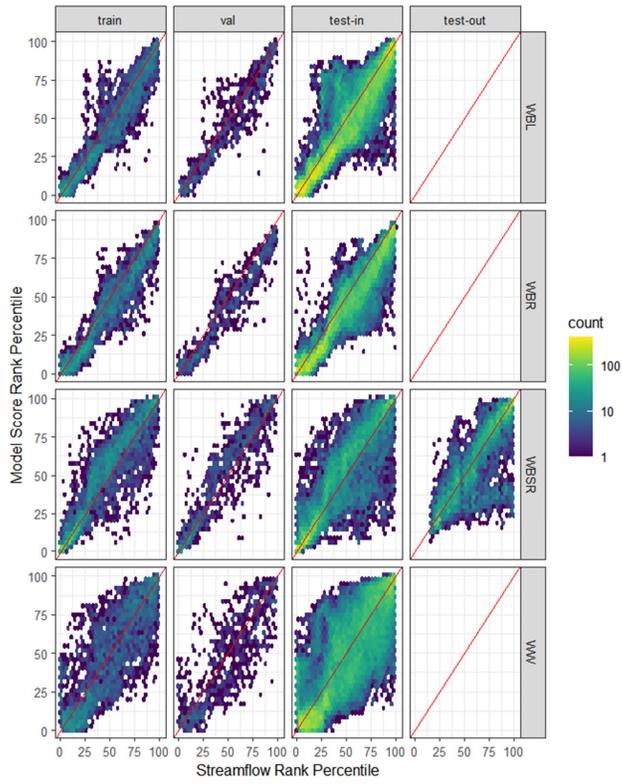


Figure S13: Predicted versus observed rank percentiles for each location and data split, shown as two-dimensional density plots. Colour indicates the number of 15-minute observations in each bin and is on a logarithmic scale. The red line is a 1:1 line. This second panel shows four of the eleven camera sites.

160

165

S4 Multi-panel figure showing view from all sites

This figure displays a representative view from all 11 camera sites (at eight locations).



Figure S14: Images captured at each site at between 7am and 8am on September 5th, 2023, during a period of low flow. The photos are provided to show general view for each location. Photo credit U.S. Geological Survey.