Canadian Science Publishing
NRC Research Press journals | Publishing Services

Canadian Journal of Fisheries and Aquatic Sciences
Journal canadien des sciences halieutiques et aquatiques

Statistical arrival models to estimate missed passage counts at fish weirs

Journal:	*Canadian Journal of Fisheries and Aquatic Sciences*
Manuscript ID:	Draft
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Sethi, Suresh; US Fish and Wildlife Service, Bradley, Catherine; US Fish and Wildlife Service, Fisheries and Ecological Services
Keyword:	fisheries management, interpolation, phenology, SALMON < Organisms, weir

Statistical arrival models to estimate missed passage counts at fish weirs

Suresh Andrew Sethi*,[1], Catherine Bradley[2]

* Corresponding author: suresh_sethi@fws.gov, telephone: +1 (907) 786-3655, fax: +1 (907) 786-3978

[1]U.S. Fish and Wildlife Service, Fisheries and Ecological Services Division, 1011 E Tudor Road, Anchorage, Alaska 99503, USA.

[2]U.S. Fish and Wildlife Service, Fisheries and Ecological Services Division, 101 12th Ave, Fairbanks, AK 99701, USA; catherine_bradley@fws.gov

Abstract

Missed counts are commonplace when enumerating fish passing a weir. Typically connect-the-dots linear interpolation is used to impute missed passage; however, this method fails to characterize uncertainty about estimates, and cannot be implemented when the tails of a run are missed. Here, we present a statistical approach to imputing missing passage at weirs which addresses these shortcomings, consisting of a parametric run curve model to describe the smoothed arrival dynamics of a fish population and a process variation model to describe the likelihood of observed data. Statistical arrival models are fit in a Bayesian framework and tested with a suite of missing data simulation trials and against a selection of Pacific Salmon (*Oncorhynchus* spp.) case studies from the Yukon River drainage, Alaska, U.S.A. When compared against linear interpolation, statistical arrival models produced equivalent or better expected accuracy and a narrower range of bias outcomes. Statistical arrival models also successfully imputed missing passage counts for scenarios where the tails of a run were missed.

Introduction

Fish which migrate en masse along river systems during part of their life history provide opportunity to census populations by directly counting passage at a weir. Weirs are relied upon heavily in harvested salmonid systems to provide information for run reconstruction, with weir sites typically sighted in upper reaches of river networks and providing estimates of adult spawning escapement after passing through a harvest gauntlet down river (e.g. Fleischman et al. 2013). While typically used to assess adult fish migrating into river systems, they can also be used to assess the abundance of out-migrating juvenile fish such as Pacific Salmon smolts (*Oncorhynchus* spp.; Bradford 1999).

In theory, weirs produce complete enumeration of a fish run passing a point in space, providing a gold standard for accurate enumeration data. In practice, field crews contend with myriad challenges in successfully installing and operating a weir—high water can blow a weir out or make it unsafe to operate, woody debris can overwhelm weir structures, river ice can delay the installation of weirs or require their early removal, and bears can vandalize weirs or harass weir operators (Figure 1). Thus, missing data at weirs are a common occurrence leading to partial enumeration of fish populations.

Attempts to impute missing data at weirs typically involve a "connect-the-dots" linear interpolation scheme (e.g. Gewin et al. 2005; Johnson et al. 2007). This approach is attractive in that it is simple to implement; however, linear interpolation requires a passage observation before and after a missing datum and thus is not an option to impute missing data in the tails of a run when weirs are installed late or pulled early. Furthermore, linear interpolation does not produce estimates of uncertainty for imputed missing passage counts, potentially communicating false precision to users of the data.

An alternative to linear interpolation is to take a model-based approach and implement a statistical framework with which to make inference about passage counts at a weir. Several related efforts have been developed for estimating abundance of fish in streams associated with periodic ground-based stream surveys (e.g. Hilborn et al. 1999; Su et al. 2001). These applications explicitly model the arrival of fish in streams and the "decay" dynamics as fish exit a stream (i.e., become unavailable for detection) through depredation by predators, decomposition, or emigration. The decay model is necessary to avoid double counting of previously observed fish. Subsequently, a probability model is asserted to describe variability of observations about the underlying arrival and decay models describing fish presence in the stream, providing a likelihood for observed fish counts. Total stream abundance enters the model framework as a scalar on the arrival and decay processes and is estimated using the likelihood for the observed data. Hilborn et al. (1999) implemented the approach using maximum likelihood estimation, and Su et al. (2001) extended the approach into a hierarchical Bayesian framework allowing for analysis of multiple years of data in manner which captures across-year correlation in stream arrival and decay processes. While these approaches deal with stream surveys on Pacific Salmon with an additional complexity of needing to correct observations for the fact that fish persist in stream for some variable amount of time, much of the underlying approach is applicable to imputation of missing data at weirs.

Conceptually, weir-based fish counts are simpler than ground-based stream surveys because weir passage is a unidirectional, one-time event. As such, an arrival model can be specified to describe the passage of a run of fish at the weir; however, a decay model which corrects for potential double counting of previously observed fish is unnecessary. Once fit, the arrival model and assessment of variability of realized passage counts about the smooth

underlying arrival model can be used to estimate passage counts for missing dates. For example, the arrival model of fish at a weir could be specified as following a cumulative Normal distribution-shaped curve, and the deviations about the smooth arrival model specifying actual weir passages could be modeled as Negative Binomial random variables. Subsequently, the Negative Binomial likelihood model can be used to inform best fitting parameters describing the arrival curve for a dataset of daily passage counts, and ultimately make predictions about missing passage dates. Because the approach is based upon a statistical framework, uncertainty about missing passage estimates can be assessed, for example by providing confidence (or credibility) intervals for missed passage estimates.

In this article, we introduce a parametric statistical approach based upon fitted run arrival models in order to impute missing data and estimate total passage at weirs. Models are fit in a Bayesian framework, providing a straightforward means to summarize uncertainty about total run size estimates, arrival model characteristics (e.g. peak run date), and estimates of predicted passage counts on missing-observation dates. Subsequently, information-theoretic model selection is used to assess strength of evidence for different proposed arrival models and observation processes. Estimator performance is tested by simulating weir data modeled after Pacific Salmon run dynamics under a suite of missing data scenarios including missing tails, missing peak run days, and data collection efforts of only 5 days a week (the "take weekends off" scenario). Performance of parametric run models is compared against a simple linear interpolation, "connect-the-dots" scheme. Finally, the parametric run curve approach is illustrated on observed weir data for Pacific Salmon runs in the Yukon River, Alaska, drainage.

Results indicate that DIC-preferred statistical arrival models have lower or equivalent average bias when compared against the standard linear interpolation approach to missing data at

weirs in all test scenarios, and produce a considerably narrower range in bias outcomes than the simple linear interpolation approach. Statistical arrival models can be successfully applied to particularly ill-behaved datasets (e.g. missing observations from the tails of the run curve) where linear interpolation can fail dramatically. Furthermore, parametric models provide statistical estimates of arrival dynamics parameters informative for assessing fish phenology.

Methods

The statistical framework implemented to estimate missing data at weirs requires specification of two key processes: *i*) an underlying run curve model that governs fish arrival dynamics at the weir, and *ii*) an error model that governs the noise about the smooth arrival model exhibited by a realized fish run and which serves as the likelihood model for the observed weir counts. Subsequently during model fitting, the parameters specifying the shape of the arrival model, a total run size scalar for the arrival model (see below), and the parameters specifying the error process model are estimated. Statistical arrival models are implemented in a Bayesian framework, providing a means to directly estimate missing data while accounting for uncertainty by treating missing passage dates as derived parameters for which posterior summaries are generated. The key parameter of interest, total run size, is defined as a derived parameter equal to the sum of all estimated missing passage dates plus the known (i.e. observed) passage counts.

Statistical models

Passage dynamics at the weir are specified by a run arrival curve model that describes the cumulative proportion, p_t, of the run that has passed the weir at time step, t. In what follows, we

define a time step as a 24 hour day; any choice of time step may be implemented, however, passage counts need match the specified time step and daily counts are the most commonly recorded weir data. We implemented two candidate arrival models: a run curve described by the function for a Normal cumulative distribution,

$$p_t = F_N(t \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{(v-\mu)^2}{2\sigma^2}} dv \qquad \text{eq. 1,}$$

and a run curve described by the function for a skew-Normal cumulative distribution (Azzalini 1985),

$$p_t = F_{SN}(t \mid \xi, \omega, \alpha) = \frac{2}{\omega} \phi\left(\left.(t-\xi)\right/_{\omega}\right) \Phi\left(\alpha\left.(t-\xi)\right/_{\omega}\right) \qquad \text{eq. 2,}$$

where F_N and F_{SN} are notation for the Normal and skew-Normal cumulative distribution functions, respectively, μ is the location parameter and σ is the scale parameter for the Normal distribution arrival model, ξ is the location parameter, ω is the scale parameter, α is the shape parameter (describing skewness) for the skew-Normal distribution, ϕ is the standard normal density function, and Φ is the standard Normal cumulative distribution function. Subsequently, daily passage counts under the arrival model, c_t, are calculated by taking the proportion of the run that arrives over a preceding time step multiplied by a total run size scalar, S:

$$c_t = S \times (p_t - p_{t-1}) \qquad \text{eq.3.}$$

To avoid confusion, it is worth emphasizing that the cumulative distribution functions used to specify arrival dynamics do not represent probability models for weir passage counts; they are merely convenient mathematical functions to describe the shape of the arrival curve of fish at a weir.

Next, a likelihood for the observed passage counts is specified by asserting a probability model for the process variation about passage counts predicted under the arrival model. We considered both a Normal and Negative Binomial probability model for an observed passage count, c_t^o :

$$c_t^o \sim Normal(c_t, \tau) \quad \text{eq. 4,}$$

or

$$c_t^o \sim NegativeBinomial\left(\lambda = {}^{\theta}/_{(\theta + c_t)}, \theta\right) \quad \text{eq.5,}$$

where the specified parameterization of the success probability parameter, λ, in the Negative Binomial model produces an expectation of c_t^o (i.e. observed passage) equal to c_t (i.e. passage under the smooth arrival curve) as a function of the dispersion parameter θ (e.g. Ntzoufras 2009). The Normal process variation model implies that variability in realized passages about the smooth underlying arrival curve is constant across the run, whereas the Negative Binomial model assumes that passage count variability scales positively with the magnitude of passage. Depending on the specified arrival model and error model combination, likelihoods of observed counts are conditional upon four or five parameters:

$$L_{N-N}(\boldsymbol{c}^o \mid \mu, \sigma, S, \tau) = \prod_{t \in T} \left[\frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(c_t^o - c_t)^2}{2\tau^2}\right) \right] \quad \text{eq. 6,}$$

$$L_{SN-N}(\boldsymbol{c}^o \mid \xi, \omega, \alpha, S, \tau) = \prod_{t \in T} \left[\frac{1}{\tau\sqrt{2\pi}} \exp\left(-\frac{(c_t^o - c_t)^2}{2\tau^2}\right) \right] \quad \text{eq. 7,}$$

$$L_{N-NB}(\boldsymbol{c}^o \mid \mu, \sigma, S, \theta) = \prod_{t \in T} \left[\binom{c_t^o + \theta - 1}{c_t^o} (\lambda)^\theta (1 - \lambda)^{c_t^o} \right] \quad \text{eq. 8,}$$

$$L_{SN-NB}(\boldsymbol{c}^o \mid \xi, \omega, \alpha, S, \theta) = \prod_{t \in T} \left[\binom{c_t^o + \theta - 1}{c_t^o} (\lambda)^\theta (1 - \lambda)^{c_t^o} \right] \quad \text{eq. 9,}$$

where $\boldsymbol{c}^o$ indicates a vector of observed counts on dates contained in the set of observed dates T, and subscripts *N—N, SN—N, N—NB, SN—NB* indicate Normal arrival curve – Normal deviates, skew-Normal arrival – Normal deviates, Normal arrival-Negative Binomial deviates, and skew-Normal arrival-Negative Binomial deviates, respectively. In a Bayesian framework, the above likelihood equations are then used to compute posterior probabilities for the set of fitted and derived parameters in a given model. In what follows, estimates of fitted and derived parameters are indicated by the "^" accent.

Model fitting

Arrival models were fit in a Bayesian framework using Markov Chain Monte Carlo algorithms as implemented in the WinBUGS software platform (Lunn et al. 2000) and run from the R statistical programming environment (RDCT 2014) using the package R2WinBUGS (Sturtz et al. 2005). Example code is provided in Supplement 1 of the Supplementary Materials. Diffuse priors were specified for fitted parameters (Table 1). Input data to model fits include a

vector of sample dates with an associated vector of observed passage counts, and a vector of missing dates for which passage estimates are produced. Dates are treated as calendar days and thus range from 1 to 365 (i.e. Julian day). WinBUGS treats missing dates as parameters to be estimated and thus returns a posterior distribution for the passage estimate for each missing date. The sum of all observed passages and of all estimated passages for missing dates was treated as a derived parameter, returning a posterior distribution for the estimated total run size. Because observed passage counts are known quantities, this estimate of total run only incorporates uncertainty about estimated missing passage counts.

WinBUGS does not have a native skew-Normal distribution function. Thus, we implemented a numerical integration routine within WinBUGS models to approximate the skew-Normal cumulative distribution function from the skew-Normal density function using standard normal density and cumulative distributions following equation 2 (Supplementary Materials, Supplement 1).

To improve model fitting, we included additional prior information when fitting statistical arrival models by including two dates approximately one month before the start of the run and two dates approximately one month after the end of the run with zero passage counts. Preliminary analyses indicated that inclusion of known-zero passage dates improved estimation of arrival models, particularly when observed data was missing for the tails of the run as may occur when a weir begins operation after the commencement of a run or is pulled before the close of a run. The timing of these dates were selected to be far enough before and after the run such that there is high confidence they represent true zero passage dates; these represent a form of prior knowledge because a priori, the timing of the fish run to be modeled isn't known, however, analysts typically have enough information about the timing of a run to propose dates

early enough and late enough to be true zero passage days. We conducted simulations to test the sensitivity of estimation results to the number and placement of known zero passage dates and found total run size estimates to be robust (Supplementary Materials, Supplement 2).

Performance testing

We compared the performance of the statistical models against a suite of trials in which we simulated fish runs modeled after Pacific Salmon data from the Yukon River. The basic approach to performance testing involved three phases: *i*) assert "true" fish run arrival dynamics and simulate realized fish runs, *ii*) implement observed fish runs which include missing passage counts (e.g. the initial 15% of the run went unobserved), and *iii*) estimate missing passage counts and summarize performance.

Simulation trials included several plausible scenarios weir operations might experience in the field (Table 2), including scenarios where the peak of the run was missed (+/- 2 days about peak run day), scenarios where tails of the run were missed (initial 15% of run missed), and a scenario of sampling five days a week with weekends off. A selection of Normal and skew-Normal arrival dynamics were simulated.

Simulated daily and total passage counts were calculated by asserting a smooth arrival process and then adding process variation noise, following the approach outlined above: latent daily passage counts are calculated as the proportion of an asserted arrival curve that passes on a given date multiplied by a total run size scalar, and subsequently, realized daily passages are drawn from probability distributions with expectations equivalent to the latent passage count. We fixed the total run size scalar at 10000 fish for each simulated data set, but because realized true daily passages included process noise, realized total run size typically differed from 10000.

Accordingly, subsequent performance of estimators for run size with missing passage counts was assessed relative to the realized true run size for a given simulated data set.

Salmon data from Yukon River weirs suggests that daily passage variability scales positively with passage magnitude (Figure 2), therefore simulated data scenarios incorporated Negative Binomial process variation about a given arrival model. The amount of process variation specified for simulated fish runs was chosen to reflect variation observed in actual Yukon River Pacific Salmon data, and was fixed at $\theta = 7.0$. For each simulated fish run, the parameters describing the arrival model functions were randomly drawn from probability distributions, such that each simulated fish run had a unique shape and timing. Location parameters for both Normal and skew-Normal arrival models were drawn from a Normal distribution with a mean of day 185 (i.e. Julian date corresponding to July 4 in non leap years) and standard deviation of 2.5 days. Scale parameters for both Normal and skew-Normal arrival models were drawn from a log-Normal distribution with a mean of $\ln(7.5)$ and standard deviation of $\ln(1.1)$. Skewness aside, these distributions for fish run arrival timing (location parameters) and fish run compression (scale parameters) result in runs of approximately 40 days in length with a peak around July 4. Finally, for simulation scenarios with skew-Normal arrival curves, the shape parameter describing skewness was fixed at a magnitude of 2.0 with the direction of skew randomly chosen for each simulated run (i.e. $\alpha \in \{-2.0, 2.0\}$), providing a moderate amount of skewness to simulated fish runs (Figure2c). Note, because the direction of skew was randomly chosen, simulation scenarios representing late-entry of weirs where the initial 15% of the run was missed are implemented for both left- and right-skew runs.

For each simulated data set, we fit four statistical models —a Normal arrival curve with Normal deviates, a Normal arrival curve with Negative Binomial deviates, a skew-Normal

arrival curve with Normal deviates, and a skew-Normal arrival curve with Negative Binomial deviates—to the simulated observed data and estimated missing data following Bayesian implementation detailed above. Preliminary model analyses indicated Markov chains converged quickly. Subsequently, two parallel chains each with a 5000 iteration burn in period, a thin rate of 10 iterations, and 1000 posterior parameter draws stored were implemented for each model fit. Model convergence was assessed using the Gelman-Brooks-Rubin $\hat{R}$ statistic (Brooks and Gelman 1998) for the key parameter of interest, derived total run size estimate (i.e. observed passages plus the sum of imputed missing passage counts).

In addition to statistical arrival models, we implemented a "connect the dots" linear interpolation estimator commonly used to impute missing data at weirs. The "connect the dots" approach is attractive because of its simplicity to implement, however the approach does not characterize the uncertainty about estimates of missing passage counts. With this estimator, missing passage days are estimated using the slope, m, of a straight line between the observed passage count days nearest to a missing day or sequence of missing days:

$$\hat{m}_{t_a,t_b} = \left(c^o_{t_b} - c^o_{t_a}\right) \Big/ |t_b - t_a| \qquad \text{eq. 10,}$$

$$\hat{c}^o_t = c^o_{t_a} + (t - t_a)\,\hat{m}_{t_a,t_b} \qquad \text{eq. 11,}$$

where t_a is the nearest date with an observed passage count preceding a given missing-observation passage day, and t_b the nearest date with an observed passage count following a given missing-observation passage day. The "connect the dots" estimator requires dates with observed passage counts both before and after a given missing date, and thus cannot directly

interpolate missing tails of the run without additional assumptions. We were unable to identify a standard approach to linearly impute missing data in the tails of runs from published literature. Therefore, we implemented an approach whereby the slope parameter for linear imputation is estimated by drawing a straight line from the maximum observed passage day (as an empirical estimate of the peak of the run) to the first observed passage day in cases when the initial tail of the run is missing, or to the last observed passage day in cases where the terminal tail of the fish run is missing. Subsequently, cumulative passage in the missing tail, $\hat{c}^o_{tail}$, is estimated by taking the area of the triangle formed by extrapolating the estimated slope to a point of zero passage before (after) the first (last) observed passage date when the initial (terminal) tail of the run is missing:

$$\hat{m}_{tail} = {(c^o_{max} - c^o_{t^*})}\Big/{|t_{max} - t^*|} \quad \text{eq. 12,}$$

$$\hat{c}^o_{tail} = {(c^o_{t^*})^2}\Big/{2\hat{m}_{tail}} \quad \text{eq. 13,}$$

where c^o_{max} is the maximum observed passage count which occurs on date t_{max}, and $c^o_{t^*}$ is the first (last) observed passage count on date t^* when cumulative passage in the initial (terminal) tail of the run is to be estimated.

A total of 150 data sets were generated for each simulation scenario (Table 2). Simulated datasets for which one or more of the four fitted statistical models resulted in $\hat{R} > 1.2$ for the total run size estimate, or for which the linear interpolation estimator failed (see below), were dropped from consideration. Estimator performance was measured as average percent bias, interquartile range in percent bias, and root mean squared error. In addition, for statistical

299 estimators implemented in a Bayesian approach, we report the median 95% highest posterior
300 density intervals (HPDI) for the derived estimate of the total run. Finally, we used DIC-based
301 model selection (Spiegelhalter et al. 2002; Celeux et al. 2006) to assess whether the "correct"
302 statistical model which matched the simulated arrival and process noise model was identified as
303 the best fitting model. We calculated the model complexity penalty in DIC using one half the
304 variance of the posterior deviance, following an approach outlined in Gelman et al. (2004;
305 Section 6.7).
306
307 *Case studies*
308 The suite of four statistical estimators and the linear imputation approach to estimating
309 missing data at weirs were tested against a selection of Pacific Salmon data sets with missing
310 passage dates from resistance board weirs operated by the U.S. Fish and Wildlife Service on the
311 Gisasa and East Fork Andreafsky rivers in the Yukon River drainage, Alaska, USA, during 1998
312 and 2000 (Table 3; data publicly available from the U.S. Fish and Wildlife Service, email:
313 ak_fisheries@fws.gov). Yukon River salmon were selected as a study system because weir
314 operations in this high latitude, large river system often experience challenges with blowouts
315 from high water and difficulty in installing (maintaining) weirs early (late) in the season due to
316 ice. Furthermore, salmon in these systems follow the typical unimodal arrival curve for
317 migrating Pacific Salmon (Figure 2). Case study datasets were chosen to represent a selection of
318 missing data scenarios, run sizes, and Pacific Salmon species (Chinook: *O. tshawytscha*; Chum:
319 *O. keta*; Coho: *O. kisutch*; Pink: *O. gorbuscha*) datasets were selected prior to completion of
320 simulation analyses. *A priori* placement of zero passage days (Table 3) were made based on the
321 shape of the arrival curve or, in the instance where one or both tails were unobserved, by

referencing data collected in previous years from a given system. Sensitivity analyses indicated total run size estimates for case studies were robust to zero passage day placements (Supplementary Materials, Supplement 2). Two parallel chains each with a 7500 iteration burn in period, a thin rate of 10 iterations, and 1500 posterior parameter draws stored were implemented for model fits. The same uninformative priors and convergence diagnostics used to fit statistical arrival models for data simulations (above) were also used for case study data (Table 1).

Results

Simulations

Of the 3000 statistical models fit to simulated data (5 scenarios, 150 datasets per scenario, 4 statistical models per dataset), 50 fits failed to converge based upon the chain mixing criteria of $\hat{R} > 1.2$ for the derived total run size estimate, resulting in the removal of 48 datasets from subsequent performance testing. Convergence failure was most common for simulation scenarios with the initial 15% of the run missing (56% of failures). Amongst statistical estimators, most convergence failures were attributable to *SN-N* model (58% of converge failures) and the *SN-NB* model (40% of convergence failures). An additional two datasets were removed from further analysis because a linear interpolation estimate could not be calculated. In both cases, the initial date for which passage was observed, t^*, also corresponded to the maximum passage count observed, t_{max}, resulting in an undefined slope and inability to extrapolate missing passage dates in the tail of the simulated runs.

Total run size estimation performance varied across models and simulation scenarios; however, in most cases missing passage estimates could be imputed with accuracy and

reasonable precision. The 'Weekends Off' scenario resulted in accurate and precise missing passage imputation estimates across all estimators, with a maximum average bias of only 0.4% of true run size for the linear interpolation estimator, and minimum average bias of 0.004% attributable to the *N-N* statistical model (Table 4; "Weekends off (Normal)" scenario). Statistical estimators indicated precise total run size estimates with 95% HPDI half widths on the order of ± 14-17% of total missed passage (± 4-5% of total run size). Seventy percent of the datasets were best described by the *N-NB* statistical arrival model as measured by DIC; however, the average difference in DIC between the *N-NB* model and the *SN-NB* model was 0.04 DIC units, indicating that the two models were comparably supported by the data (data not shown but available from the authors).

Imputation of missing passage counts on scenarios with missed peak run days was generally successful for both the Normal and skew-Normal arrival curve data generating processes simulated. Accuracy and precision of total run estimates were comparable for data simulated under the Normal and skew-Normal arrival models. For example, expected bias of the DIC-preferred model (*N-NB*) under the Normal arrival curve scenario was -0.3% with a HPDI half-width of approximately ± 18% of total missed passage, whereas the DIC-preferred model (S*N-NB*) under the skew-Normal arrival curve scenario resulted in slightly larger expected bias of -2.6% but a slightly tighter HPDI half-width of approximately ± 15% of total missed passage (Table 4; "±2 days about the peak missed" scenarios). In both scenarios, DIC-based model selection correctly identified the correct approximating model the majority of time. The *SN-NB* statistical model performed well in both missing peak days scenarios, evincing equivalent performance in terms of expected bias, range in bias, and RMSE as the *N-NB* model in scenarios with a Normal arrival curve data generating process, and top performance in missing peak days

scenarios with a skew-Normal arrival curve data generating process. Linear interpolation imputation estimators had performance within the range observed for statistical arrival curve estimators; however, the former was never the best estimator in terms of expected bias, range in bias, and RMSE.

Simulation scenarios for which the initial 15% of the run was missed—representing, for example, late installation of a weir in the field—presented a greater range in estimator performance and lower precision when compared against scenarios with missing passage dates within the interior of arrival curves. Accuracy was generally higher for missing tail scenarios with Normal arrival dynamics, resulting in expected biases ranging from 0.5 to 11.4% across estimators, as compared to missing tail scenarios with skew-Normal arrival dynamics, which resulted in expected biases ranging from 6.3 to 17.4 %. DIC-based model selection performed well for the Normal arrival curve scenario, correctly identifying the *N-NB* model as the best fit model in over 80% of trials (Table 4; "First 15% missed (Normal)" scenario); however, DIC-based model selection in the case of skew-Normal arrival dynamics indicated ambiguity, supporting both the *N-NB* and *SN-NB* as preferred models (Table 4; "First 15% missed (skew-Normal)" scenario). Furthermore, the *N-N* model under the skew-Normal arrival scenario was never DIC-preferred but was of comparable performance to the *SN-NB* model, with average percent bias of 6.35% and 6.27%, respectively, and interquartile range of bias of 12.0% and 15.9%, respectively. Because of the high performance of the *SN-NB* model for missing-tail data scenarios for both Normal and skew-Normal arrival dynamics, but potential ambiguity in DIC-based model selection in identifying the best approximating model, we recommend implementing the *SN-NB* model when handling real data with missed passage counts for tails of unimodal-shaped runs such as for Pacific Salmon enumeration as examined herein. Finally,

linear interpolation estimators produced average percent bias comparable to the best fit statistical models for missing tail scenarios; however, linear interpolation could result in very large positive bias outcomes with maximum percent bias of 176.6% in the case of Normal arrival dynamics and maximum bias of 930% in the case of skew-Normal dynamics (Table 4).

While not a focus of the present simulation trials, statistical arrival models also produced good performance in reconstructing underlying run curve characteristics (see summary of full results in Supplementary Materials, Table S1). Arrival curve location and spread parameters were estimated with high accuracy, with expected bias for peak run date estimated to within ± 1 day and spread parameters to within ± 5% of the true estimate for DIC-preferred models. Arrival curve skew, which was constrained to ±2.0 during simulations, was accurately estimated with average absolute bias of 0.02 for the missing peak days scenario with skew-Normal arrival; however, skew estimation for the missing 15% initial tail with skew-Normal arrival was less accurate and resulted in average absolute bias of 1.92 for skew parameter estimates across simulation trials. Thus, we caution due diligence to ensure reasonable skewness estimates when implementing statistical arrival models for weir data with missing tails, for example by visualizing data and ensuring plausible fitted arrival dynamics (cf. Figure 2).

Case Studies

The *SN-NB* was the DIC-preferred model in each case study, indicating substantial left and right skew in 4 of 5 of the case studies (Figure 2d-h). Based upon total run size from the *SN-NB* statistical arrival model, imputed passage counts ranged from 11-31% of total run (Table 5). Precision of total run estimates varied, and was not strongly related to the amount of missing passage imputed or the shape of the run. For example, the *SN-NB* statistical arrival estimator

414 indicated 26% of the Chum salmon run on the Gisasa River in 1998 went unobserved, producing
415 a precise total run estimate with a 95% HPDI half width of ±5.4% of the estimated run size
416 (±21% of the total imputed missing passage). In contrast, data from the 2000 East Fork
417 Andreafsky Chum salmon run indicated 19% of the run went unobserved, producing a less
418 precise total run size estimate with a 95% HPDI half width of ±17.1% of the estimated run size
419 (±92% of the total imputed missing passage). The lowest precision total run estimate occurred
420 with the 1998 East Fork Andreafsky Coho Salmon run (±34.8% of the total run estimate; ±112%
421 of the total imputed missing passage), which exhibited high passage count variability, strong
422 skewness, and missing data in a tail of the run (Figure 2f).
423 The statistical arrival model total run estimates and those from linear interpolation agreed
424 closely for the case study runs that exhibited relatively smooth and symmetric arrival dynamics
425 (1998 Gisasa Chum and 1998 Gisasa Chinook; Figure 2d-e; Table 5); however, the linear
426 interpolator agreed less closely in cases with missing data near peak runs with high passage
427 variability. For example, the maximum observed count for 1998 East Fork Andreafsky Coho
428 (1,104 fish in a run where the total observed count equaled 5,417 fish) was followed by an
429 observed 0 passage day and subsequently observations ceased. Owing to a lack of observations
430 in the remainder of the run, linear imputation for the tail based upon the ad hoc rule outlined
431 above resulted in a 0 fish estimate and substantial underestimate of the run size relative to the
432 *SN-NB* statistical arrival model estimate (Table 5).
433
434 **Discussion**
435 Simulation results demonstrated that in many cases missing data at weirs can be
436 accurately and precisely imputed for unimodal fish runs. For example, under the simulated

process variation modeled after Pacific Salmon runs in Alaska, a weekends-off monitoring schedule (a repeated schedule of 5 days of monitoring followed by 2 days of missing data) produced little loss in accuracy for estimated total run size, while still allowing for estimation of phenological run characteristics such as peak run date.

DIC-based model selection generally performed well in identifying candidate models which reflect the asserted data generating process during simulations, although at the cost of needing to specify suites of candidate models. DIC-based model selection preferred the Negative Binomial process variation models, which reflects the error model asserted for simulated data. Within the set of arrival models with Negative Binomial process variation, DIC model selection indicated support for the skew-Normal arrival model across all scenarios, though not always as the majority DIC-preferred model. Based upon expected percent bias, range in bias outcomes, and RMSE, the *SN-NB* statistical arrival model performed equivalently or better to the *N-NB* models, and was also always DIC-preferred when faced with case-study data. Furthermore, the *SN-NB* arrival model – process variation model combination was the most general model estimated, requiring only one additional parameter than models with Normal arrival curves. The skew-Normal arrival model reduces to a Normal arrival model under a shape parameter of zero. Furthermore, the Negative Binomial process model allows for overdispersion in count models, but can also reduce to a pure Poisson count process model if the overdispersion parameter is zero. Thus, while we recommend fitting a suite of plausible statistical arrival models and subsequent DIC-based model selection, the *SN-NB* model may be a reasonable choice to apply to real data for those wishing to avoid fitting suites of models.

Statistical arrival model estimators outperformed linear interpolation in most simulation trials; however, the performance gap was not always great. On average, linear interpolation was

460 of comparable accuracy in imputing missed passage counts as the DIC-best statistical arrival
461 models, with slightly poorer bias performance when observations were missed on peak passage
462 days (Table 4). The proposed rule of thumb for implementing linear interpolation when the tails
463 of a fish run were missed performed well in many simulation trials and yielded total run
464 estimates quite close to model-based estimates in many cases; however this estimator led to a
465 wider range in bias outcomes and occasionally led to very poor estimation outcomes across
466 simulation trials (Table 4).

467 The comparable performance of linear interpolation against fitted arrival models begs the
468 question as to whether the statistical complication is worth the trouble when imputing missing
469 passages at weirs? For simple applications with only a few missing passage counts occurring at
470 non-peak passage dates, linear interpolation is likely sufficient, producing an accurate estimate of
471 total run size. The model-based framework implemented in a Bayesian framework did, however,
472 present several advantages over linear interpolation. First, uncertainty estimates about missing
473 passages were produced, providing transparency about the precision of total run size. For
474 example, the total run size estimates from the linear interpolation estimator (7179) and the *SN-*
475 *NB* statistical arrival model (7876) for the 1998 East Fork Andreafsky Coho Salmon were
476 similar, differing by 697 fish; however, the statistical arrival model fit indicates considerable
477 uncertainty with a total run size 95% HPDI ranging from 6497 to 11990 fish (Table 5). Second,
478 Bayesian implementation allowed for inclusion of prior information into model estimates in a
479 unified framework, which may facilitate estimates for scenarios with substantial missing-data
480 (although there is a limit to what can be asked of a dataset). Finally, the model-based framework
481 provides statistical estimates of biologically relevant information on run arrival dynamics, where
482 the median of the fitted arrival models represents the date at which 50% of the run has passed the

weir, the mode represents the peak run day after which fishery managers can expect declining passage counts, and the spread of the arrival model provides information about fish run compression. For example, Sethi and Tanner (2014) characterized annual variability in run dynamics of Chinook Salmon on the Togiak River in the Bristol Bay region of Alaska, by comparing estimates of peak run date and run compression at a resistor board weir across a suite of years using methods similar to the statistical arrival model approach outlined above.

While data simulations and case studies herein focused on missed passage counts through weirs, the statistical arrival model approach to imputation could also be applied to other fish enumeration techniques which produce counts that follow the underlying arrival dynamics of a target fish population. Example fish counting projects include counting towers (Anderson 2000), hydroacoustic enumeration (e.g. Enzenhofer et al. 1998), or test fisheries (Flynn and Hilborn 2004). A key requirement for such applications is that passage counts need either completely enumerate all fish present on observation days, or need represent a consistent index of passage on observation days (e.g. with constant or controllable "catchability").

Acknowledgements

We thank R. Brown (USFWS) and A. Martin (USFWS) for assistance with data for this project. R. Brown also contributed to the conception of this project. We thank J. Reynolds (USFWS) for comments which improved an earlier draft of this piece. The findings and conclusions in this article are those of the authors and do not necessarily represent the views of the U.S. Government.

References

Anderson, C. J. 2000. Counting tower projects in the Bristol Bay area, 1955–1999. Alaska Department of Fish and Game, Division of Commercial Fisheries Regional Information Report No. 2A00–08. Anchorage, Alaska.

Azzalini, A. 1985. A class of distributions which includes the Normal ones. Scand. J. Stat. **12**: 171-178.

Bradford, M.J. 1999 Temporal and spatial trends in the abundance of Coho salmon smolts from Western North America. T. Am. Fish. Soc. **128**: 840-846.

Brooks, S.P., and Gelman, A. 1998. General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**: 434–455.

Celeux, G., Forbes, F., Robert, C.P., and Titterington, D.M. 2006. Deviance information criteria for missing data models. Bayesian Anal. **1**: 651–674.

Enzenhofer, H.J., Olsen, N., and Mulligan, T.J. 1998. Fixed-location hydroacoustics as a method of enumerating migrating adult Pacific salmon: comparison of split-beam acoustics vs. visual counting. Aquat. Living Resour. **11**: 61-74.

Fleischman, S.J., Catalano, M.J., Clark, R.A., and Bernard, D.R. 2013. An age-structured state-space stock–recruit model for Pacific Salmon (*Oncorhynchus* spp.). Can. J. Fish. Aquat. Sci. **70**: 401-414.

Flynn, L., and Hilborn, R. 2004. Test fishery indices for sockeye salmon (*Oncorhynchus nerka*) as affected by age. Can. J. Fish. Aquat. Sci. **61**: 80-92.
composition and environmental variables

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. 2004. Bayesian data analysis. Chapman and Hall, Boca Raton, FL, USA.

Gewin, C.S., and VanHatten, G.K. 2005. Abundance and run timing of adult Pacific Salmon in the East Fork Andreafsky River, Yukon Delta National Wildlife Refuge, Alaska, 2003. U.S. Fish and Wildlife Service Data Series Report 2005-10, Anchorage, Alaska.

Hilborn, R., Bue, B.G., and Sharr, S. 1999. Estimating spawning escapement from periodic counts: a comparison of methods. Can. J. Fish. Aquat. Sci. **56**: 888-896.

Johnson, D.H., Shrier, B.M., O'Neal, J.S., Knutzen, J.A., Augerot, X., O'Neil, T.A., and Pearsons, T.A. (Eds.) 2007. Salmon field protocols handbook. American Fisheries Society, Bethesda, Maryland.

549 Lunn, D.J., Thomas, A., Best, N., and Spiegelhalter, D. 2000. WinBUGS – a Bayesian modelling
550 framework: concepts, structure, and extensibility. Stat. Comput. **10**: 325–337.
551
552 Ntzoufras, I. 2009. Bayesian modelling using WinBUGS. Wiley, Hoboken, NJ, USA.
553
554 R Development Core Team (RDCT). 2014. R: a language and environment for statistical
555 computing. R Foundation for Statistical Computing, Vienna, Austria.
556
557 Sethi, S.A., and Tanner, T. 2014. Spawning distribution and abundance of a northern Chinook
558 population. Fisheries Manag. Ecol. **21**: 427-438.
559
560 Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. 2002. Bayesian measures
561 of model complexity and fit (with discussion). J. Roy. Stat. Soc. Ser. B **64**: 583–639.
562
563 Sturtz, S., Ligges, U., and Gelman, A. 2005. R2WinBUGS: a package for running WinBUGS
564 from R. J. Stat. Softw. **12**: 1–16.
565
566 Su, Z., Adkison, M.D., and Van Alen, W. 2001. A hierarchical Bayesian model for estimating
567 historical salmon escapement and escapement timing. Can. J. Fish. Aquat. Sci. **58**: 1648-1662.

Table 1. Prior specifications[a].

Parameter description	Prior
Run size scalar	$\log(S) \sim Normal(\mu_S = 7.5,\ \sigma_S = 2.0)$
Location parameter of the Normal arrival model	$\mu \sim Uniform(150, 300)$
Scale parameter of the Normal arrival model	$\sigma \sim Uniform(1,50)$
Location parameter of the skew-Normal arrival model	$\xi \sim Uniform(150, 300)$
Scale parameter of the skew-Normal arrival model	$\omega \sim Uniform(1, 50)$
Shape parameter of the skew-Normal arrival model	$\alpha \sim Uniform(-10, 10)$
Scale parameter for the Normal process error model	$\tau \sim Uniform(0.1, 1000)$
Dispersion parameter for the Negative Binomial process error model	$\theta \sim Gamma(0.1, 0.1)$

[a]Arrival models are parameterized in terms of a daily time step. For example, the prior specification for the Normal arrival model constrain the peak run date, μ, to lie between Julian day 150 and day 300.

Table 2. Simulation scenarios.

Observation scenario	Arrival model
Weekends off: observe 5 days, miss 2 days, ...	Normal
Run peak missed: a 5-day period (+/- 2 days) about the run peak is missed	Normal
Run peak missed: a 5-day period (+/- 2 days) about the run peak is missed	skew-Normal
Late installation: first 15% of arrival curve is missed	Normal
Late installation: first 15% of arrival curve is missed	skew-Normal

Table 3. Case study data summary.

Run	Observed dates (Julian day)	Description of missing data	*A priori* zero passage count placement (Julian day)
1998 Gisasa River Chum Salmon	172-214	Neither tail fully observed; sequence of missing observations preceding the run peak	158-160; 225-227
1998 Gisasa River Chinook Salmon	175-214	Sequences of missing days in both tails of the run	168-170; 225-227
1998 East Fork Andreafsky River Coho Salmon	203-255	Sequence of missed observations after run commencement; terminal tail missed	198-200; 270-272
2000 East Fork Andreafsky River Chum Salmon	182-216	Initial tail missed; sequence of missed days after the run peak	168-170; 230-232
2000 East Fork Andreafsky River Pink Salmon	182-262	Sequences of missed days after the run peak	168-170; 270-272

Table 4. Simulation trial performance summaries for statistical arrival models and linear interpolation[a].

Scenario (arrival model)	Fitted model	DIC preferred (%)	Total run size estimate percent bias				RMSE	Median total passage observed	Median total missing passage prediction	Median 95% HPDI width
			Average (%)	Minimum (%)	Maximum (%)	IQR (%)				
Weekends off (Normal)	*N-N*	0	<0.1	-11.3	9.6	5.7	435.5	7206	2872.0	820.5
	N-NB	70.0	<0.1	-10.7	10.3	5.8	426.8		2874.8	966.5
	SN-N	0	0.1	-11.0	10.3	5.6	440.7		2870.0	866
	SN-NB	30.0	0.2	-10.6	10.5	5.8	429.8		2887.0	940
	LI	.	0.4	-11.4	12.1	5.8	501.0		2872.5	.
±2 days about the peak missed (Normal)	*N-N*	0	0.6	-18.7	20.1	9.6	668.5	7395	2650.5	692.5
	N-NB	61.3	-0.3	-20.3	17.2	7.2	604.9		2542.5	943.5
	SN-N	0	0.4	-18.4	20.2	9.0	661.8		2638.5	725
	SN-NB	38.7	<0.1	-18.4	16.9	6.9	585.6		2585.5	955
	LI	.	-2.5	-24.6	20.5	11.7	885.3		2334.2	.
±2 days about the peak missed (skew-Normal)	*N-N*	0	-2.4	-29.6	42.6	15.0	1144.3	6223	3366.0	1002
	N-NB	0	-10.5	-37.6	14.1	10.7	1400.6		2703.5	1681
	SN-N	0	-1.2	-34.0	31.4	14.8	1170.4		3444.0	1077
	SN-NB	100.0	-2.6	-26.3	23.2	9.7	897.8		3406.5	1616
	LI	.	-6.5	-37.7	25.2	15.6	1339.0		2960.0	.
First 15% missed (Normal)	*N-N*	0	2.8	-10.9	50.6	8.6	978.5	8768	1344.0	2403
	N-NB	80.6	1.4	-7.5	16.2	7.3	512.8		1297.0	1809
	SN-N	0	11.4	-12.2	56.0	24.5	2043.1		2058.0	6147
	SN-NB	19.4	0.5	-10.2	14.4	7.4	544.4		1284.0	2245
	LI	.	0.5	-14.8	176.6	11.7	2335.8		526.7	.
First 15% missed (skew-Normal)	*N-N*	0	6.3	-13.4	74.3	15.9	1873.7	8715	1123.5	2306
	N-NB	54.1	12.4	-13.0	106.0	31.1	2662.7		1532.5	2902
	SN-N	0	17.4	-13.8	105.2	34.5	3016.2		2208.5	7171
	SN-NB	45.9	6.3	-13.0	172.6	12.0	1953.5		1336.5	3730.5
	LI	.	8.8	-14.1	930.0	10.8	8072.9		700.4	.

IQR= interquartile range; RMSE= root mean squared error.

[a]The true run size scalar for simulated fish runs is 10000 in all cases. Performance summaries for arrival model scenarios only include data sets for which all four statistical arrival model fits converged and for which a linear interpolation estimate could be generated (see Methods); number of datasets for arrival scenarios, top to bottom = 150, 129, 122, 150, and 149. Abbreviations: "DIC" = Deviance Information Criterion, "IQR" = Inner Quartile Range, "RMSE" = Root Mean Squared Error, "HPDI" = Highest Posterior Density Interval, "N"= Normal, "SN" = skew-Normal, "NB" = Negative Binomial, "LI" = Linear Interpolation.

Table 5. Case study results[a].

Run	Observed passage	Linear interpolation run size estimate	DIC preferred statistical model	Statistical model run size estimate	Lower 95% HPDI limit	Upper 95% HPDI limit	95% HPDI width	Statistical model imputed passage
1998 Gisasa River Chum Salmon	13225	17796	SN.NB	17770	16820	18740	1920	4545
1998 Gisasa River Chinook Salmon	1942	2361	SN.NB	2297	2246	2357	111	355
1998 East Fork Andreafsky River Coho Salmon	5417	7179	SN.NB	7876	6497	11990	5493	2459
2000 East Fork Andreafsky River Chum Salmon	22625	25089	SN.NB	27810	25040	34570	9530	5185
2000 East Fork Andreafsky River Pink Salmon	37069	43500	SN.NB	41670	39780	45150	5370	4601

[a]Abbreviations: "DIC" = Deviance Information Criterion, "HPDI" = Highest Posterior Density Interval, "SN" = skew-Normal, "NB" = Negative Binomial.

Figure Captions

Figure 1. Weir failures. Left panel: high water tops a resistor board weir on the Killey River, Alaska, Julne 2013 (credit: A. Waldo). Right panel: high water destroys a resistor board weir on the Gisasa River, Alaska, July 2014 (credit: J. Mears).

Figure 2. Simulated (a-c) and observed (d-h) Pacific Salmon runs.

Figures

Figure 1. Weir failures. Left panel: high water tops a resistor board weir on the Killey River, Alaska, Julne 2013 (credit: A. Waldo). Right panel: high water destroys a resistor board weir on the Gisasa River, Alaska, July 2014 (credit: J. Mears).

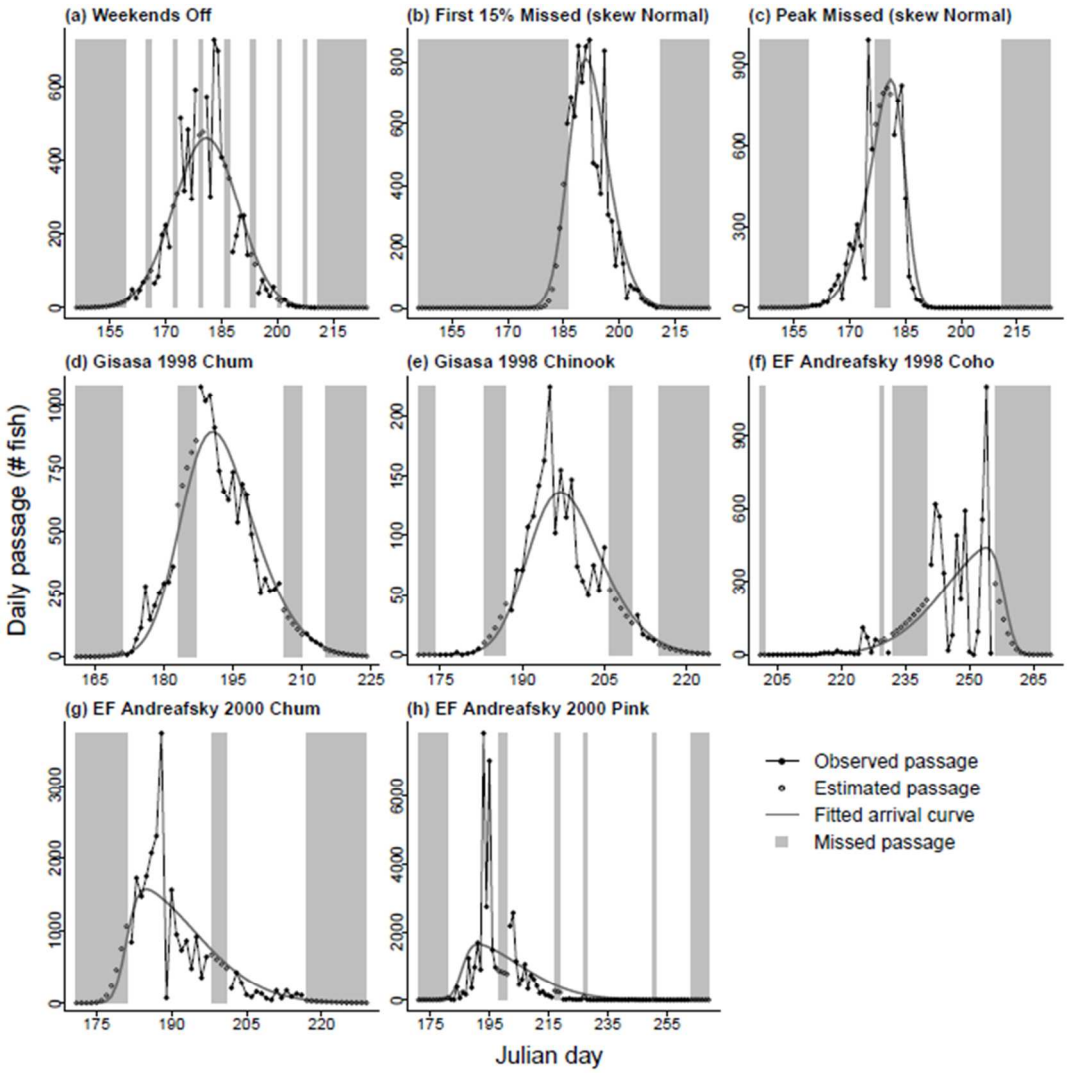
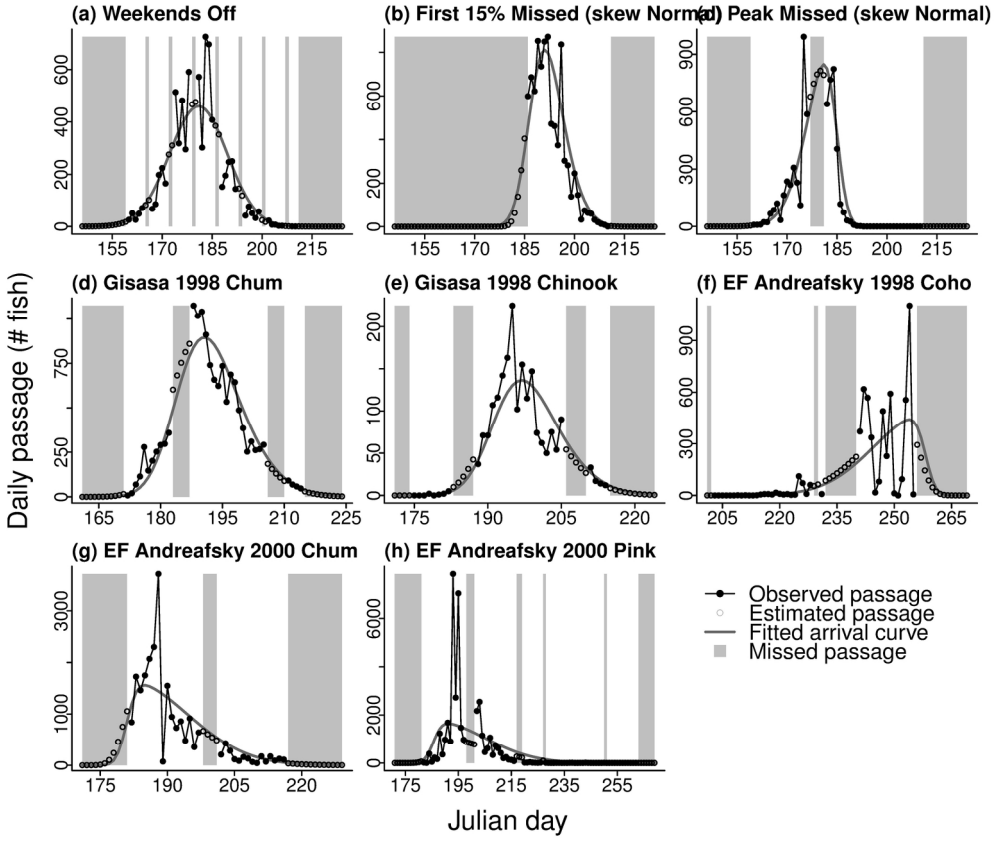


Figure 2. Simulated (a-c) and observed (d-h) Pacific Salmon runs.



179x153mm (300 x 300 DPI)

Statistical arrival models to impute missing data from fish weirs
Supplementary Materials

S.A. Sethi and C. Bradley

Contents

Supplement 1. Code supplement: WinBUGS models and example model fit from R

```
# Contents
# Part I: WinBUGS models
# Part II: Simulate data with missing passage counts and fit a SN-NB arrival model ###################################

# Part I: WinBUGS models ###########################################################
  # set a directory to store WinBUGS models (customize this to your machine)
    setwd("C:\\Users\\sureshsethi\\Desktop\\TestRuns")

  sink("NormalArrival.NormalErrorProcess.txt")
    cat("
    model {
      # Input data:
      # Date = vector of sampling dates
      # Count = vector of weir passage counts to accompany sampling dates
      # T = number of sampling occasions (i.e. length of Count vector)
      # MDate = vector of missing sampling dates to predict
      # M = number of missing sampling dates to predict
      # Parameters fit in the model
      # mu_date_run = location parameter of Normal arrival curve, equivalent to run peak (i.e. mode)
      # sigma_date_run = scale parameter of Normal arrival curve
      # sigma_fit = scale parameter for error process
      # S = the total run size scalar
      # Derived parameters estimated in the model
      # Mpred: vector of passage predictions for missing dates specified in MDate
      # RunSize: sum of predicted missing passage counts plus observed counts
      # priors
      mu_date_run ~ dunif(150,300)
      sigma_date_run ~ dunif(1,50)
      sigma_fit ~ dunif(.1,1000)
      tau_fit <- 1/(sigma_fit*sigma_fit) # error process (Normal likelihood), WinBUGS takes precision
      logS ~ dnorm(7.5 ,.25)
      log(S) <- logS # effectively flat prior for run size scalar
      # likelihood contributions
      # Step 1: calculate the estimated proportion of the run that passes in time step i
        for(i in 1:T){
          prop[i] <- (phi(((Date[i]+0)-mu_date_run)/sigma_date_run) - phi(((Date[i]-1)-mu_date_run)/sigma_date_run))
        } # end i loop
      # Step 2: calculate predicted observed passage Counts for sampled dates and their likelihood
        for(i in 1:T){
        pred[i] <- prop[i]*S
        Count[i] ~ dnorm(pred[i],tau_fit)# Normally distributed deviates
        } # end likelihood
      # derived quantities
      # predicted missing dates
        for(i in 1:M){
        Mprop[i] <- (phi(((MDate[i]+0)-mu_date_run)/sigma_date_run) - phi(((MDate[i]-1)-mu_date_run)/sigma_date_run))
        Mpred[i] <- Mprop[i]*S
        } # end i loop
      # reconstruct total run treating observed passage counts as known quantities
```

```
      RunSize <- sum(Count[])+sum(Mpred[])
    } # end model

  ",fill=T)
  sink()

sink("NormalArrival.NegativeBinomialErrorProcess.txt")
  cat("
  model {
    # Input data:
    # Date = vector of sampling dates
    # Count = vector of weir passage counts to accompany sampling dates
    # T = number of sampling occasions (i.e. length of Count vector)
    # MDate = vector of missing sampling dates to predict
    # M = number of missing sampling dates to predict
    # Parameters fit in the model
    # mu_date_run = location parameter of Normal arrival curve, equivalent to run peak (i.e. mode)
    # sigma_date_run = scale parameter of Normal arrival curve
    # theta = overdispersion parameter for error process
    # S = the total run size scalar
    # Derived parameters estimated in the model
    # Mpred: vector of passage predictions for missing dates specified in MDate
    # RunSize: sum of predicted missing passage counts plus observed counts
    # priors
    mu_date_run ~ dunif(150,300)
    sigma_date_run ~ dunif(1,50)
    logS ~ dnorm(7.5,.25)
    log(S) <- logS # effectively flat prior for run size scalar
    theta ~ dgamma(0.1,0.1) # negative binomial over dispersion parameter
    # likelihood contributions
    # Step 1: calculate the estimated proportion of the run that passes in time step i
      for(i in 1:T){
        prop[i] <- (phi(((Date[i]+0)-mu_date_run)/sigma_date_run) - phi(((Date[i]-1)-mu_date_run)/sigma_date_run))
      } # end i loop
    # Step 2: calculate predicted observed passage Counts for sampled dates and their likelihood
      for(i in 1:T){
      pred[i] <- round(prop[i]*S)
      pred.star[i] <- theta/(theta+pred[i]) #reparameterizing WinBUGS negative binomial, see main text
      temp.Count[i] <- round(Count[i])
      temp.Count[i] ~ dnegbin(pred.star[i],theta) # negative binomial likelihood
      } # end likelihood
    # derived quantities
    # predicted missing dates
      for(i in 1:M){
      Mprop[i] <- (phi(((MDate[i]+0)-mu_date_run)/sigma_date_run) - phi(((MDate[i]-1)-mu_date_run)/sigma_date_run))
      Mpred[i] <- Mprop[i]*S
      } # end i loop
    # reconstruct total run treating observed passage counts as known quantities
      RunSize <- sum(Count[])+sum(Mpred[])
    } # end model
```

```
  ",fill=T)
  sink()

sink("skewNormalArrival.NormalErrorProcess.txt")
  cat("
  model {
    # Input data:
    # Date = vector of sampling dates
    # Count = vector of weir passage counts to accompany sampling dates
    # T = number of sampling occasions (i.e. length of Count vector)
    # MDate = vector of missing sampling dates to predict
    # M = number of missing sampling dates to predict
    # invl = number of intervals to parse a time step for numerical integration by trapezoid rule
    # Parameters fit in the model
    # xi_date_run = location parameter of skewNormal arrival curve
    # omega_date_run = scale parameter of skewNormal arrival curve
    # alpha_date_run = shape parameter of skewNormal arrival curve
    # sigma_fit = scale parameter for error process
    # S = the total run size scalar
    # Derived parameters estimated in the model
    # Mpred: vector of passage predictions for missing dates specified in MDate
    # RunSize: sum of predicted missing passage counts plus observed counts
    # priors
    xi_date_run ~ dunif(150,300)
    omega_date_run ~ dunif(1,50)
    alpha_date_run ~ dunif(-10,10)
    sigma_fit ~ dunif(.1,1000)
    tau_fit <- 1/(sigma_fit*sigma_fit) # error process (Normal likelihood), WinBUGS takes precision
    logS ~ dnorm(7.5,.25)
    log(S) <- logS # effectively flat prior for run size scalar
    # likelihood contributions
    # Step 1: calculate the estimated proportion of the run that passes in time step i
      # numerical integration to approximate cdf from the pdf of skewnormal,(e.g. Azzalini 1985)
      for(i in 1:T){
      # numerical integration, intervals = invl, integration interval = 1 time unit, such that h = 1/invl
      for(j in 1:invl){
        store[j,i] <- (2/omega_date_run) * (1/sqrt(2*3.14159))*
        (
        exp(-.5* pow((((Date[i]+(j/invl))-xi_date_run)/omega_date_run),2)) *
        phi(alpha_date_run *(((Date[i]+(j/invl))-xi_date_run)/omega_date_run)) +
        exp(-.5* pow((((Date[i]+((j-1)/invl))-xi_date_run)/omega_date_run),2)) *
        phi(alpha_date_run *(((Date[i]+((j-1)/invl))-xi_date_run)/omega_date_run))
        )
        } # end j loop, sum over numerical integration vector
      prop[i] <- (.5/invl)*sum(store[,i])
      } # end i loop
    # Step 2: calculate predicted observed passage Counts for sampled dates and their likelihood
      for(i in 1:T){
      pred[i] <- prop[i]*S
      Count[i] ~ dnorm(pred[i],tau_fit)# Normally distributed deviates
      } # end likelihood
```

```
    # derived quantities
    # predicted missing dates
      for(i in 1:M){
      # numerical integration, intervals = invl, integration interval = 1 time unit, such that h = 1/invl
        for(j in 1:invl){
        Mstore[j,i] <- (2/omega_date_run) * (1/sqrt(2*3.14159))*
        (
        exp(-.5* pow((((MDate[i]+(j/invl))-xi_date_run)/omega_date_run),2)) *
        phi(alpha_date_run *(((MDate[i]+(j/invl))-xi_date_run)/omega_date_run)) +
        exp(-.5* pow((((MDate[i]+((j-1)/invl))-xi_date_run)/omega_date_run),2)) *
        phi(alpha_date_run *(((MDate[i]+((j-1)/invl))-xi_date_run)/omega_date_run))
        )
        } # end j loop, sum over numerical integration vector
        Mprop[i] <- (.5/invl)*sum(Mstore[,i])
        Mpred[i] <- Mprop[i]*S
      } # end i loop
    # reconstruct total run treating observed passage counts as known quantities
      RunSize <- sum(Count[])+sum(Mpred[])
    # reference: Azzalini, A. 1985. A class of distributions which includes the Normal ones. Scand. J. Stats. 12:171.
    } # end model

  ",fill=T)
  sink()

sink("skewNormalArrival.NegativeBinomialErrorProcess.txt")
  cat("
  model {
    # Input data:
    # Date = vector of sampling dates
    # Count = vector of weir passage counts to accompany sampling dates
    # T = number of sampling occasions (i.e. length of Count vector)
    # MDate = vector of missing sampling dates to predict
    # M = number of missing sampling dates to predict
    # invl = number of intervals to parse a time step for numerical integration by trapezoid rule
    # Parameters fit in the model
    # xi_date_run = location parameter of skewNormal arrival curve
    # omega_date_run = scale parameter of skewNormal arrival curve
    # alpha_date_run = shape parameter of skewNormal arrival curve
    # theta = overdispersion parameter for error process
    # S = the total run size scalar
    # Derived parameters estimated in the model
    # Mpred: vector of passage predictions for missing dates specified in MDate
    # RunSize: sum of predicted missing passage counts plus observed counts
    # priors
    xi_date_run ~ dunif(150,300)
    omega_date_run ~ dunif(1,50)
    alpha_date_run ~ dunif(-10,10)
    logS ~ dnorm(7.5,.25)
    log(S) <- logS # effectively flat prior for run size scalar
    theta ~ dgamma(0.1,0.1) # negative binomial over dispersion parameter
    # likelihood contributions
```

```
      # Step 1: calculate the estimated proportion of the run that passes in time step i
        # numerical integration to approximate cdf from the pdf of skewnormal,(e.g. Azzalini 1985)
        for(i in 1:T){
        # numerical integration, intervals = invl, integration interval = 1 time unit, such that h = 1/invl
        for(j in 1:invl){
          store[j,i] <- (2/omega_date_run) * (1/sqrt(2*3.14159))*
          (
          exp(-.5* pow((((Date[i]+(j/invl))-xi_date_run)/omega_date_run),2)) *
          phi(alpha_date_run *(((Date[i]+(j/invl))-xi_date_run)/omega_date_run)) +
          exp(-.5* pow((((Date[i]+((j-1)/invl))-xi_date_run)/omega_date_run),2)) *
          phi(alpha_date_run *(((Date[i]+((j-1)/invl))-xi_date_run)/omega_date_run))
          )
          } # end j loop, sum over numerical integration vector
        prop[i] <- (.5/invl)*sum(store[,i])
        } # end i loop
      # Step 2: calculate predicted observed passage Counts for sampled dates and their likelihood
        for(i in 1:T){
        pred[i] <- round(prop[i]*S)
        pred.star[i] <- theta/(theta+pred[i]) #reparameterizing WinBUGS negative binomial, see main text
        temp.Count[i] <- round(Count[i])
        temp.Count[i] ~ dnegbin(pred.star[i],theta) # negative binomial likelihood
        } # end likelihood
      # derived quantities
      # predicted missing dates
        for(i in 1:M){
          # numerical integration, intervals = invl, integration interval = 1 time unit, such that h = 1/invl
          for(j in 1:invl){
          Mstore[j,i] <- (2/omega_date_run) * (1/sqrt(2*3.14159))*
          (
          exp(-.5* pow((((MDate[i]+(j/invl))-xi_date_run)/omega_date_run),2)) *
          phi(alpha_date_run *(((MDate[i]+(j/invl))-xi_date_run)/omega_date_run)) +
          exp(-.5* pow((((MDate[i]+((j-1)/invl))-xi_date_run)/omega_date_run),2)) *
          phi(alpha_date_run *(((MDate[i]+((j-1)/invl))-xi_date_run)/omega_date_run))
          )
          } # end j loop, sum over numerical integration vector
          Mprop[i] <- (.5/invl)*sum(Mstore[,i])
          Mpred[i] <- Mprop[i]*S
        } # end i loop
      # reconstruct total run treating observed passage counts as known quantities
        RunSize <- sum(Count[])+sum(Mpred[])
      # reference: Azzalini, A. 1985. A class of distributions which includes the Normal ones. Scand. J. Stats. 12:171.
      } # end model

    ",fill=T)
    sink()
############################## End Part I ##########################################################################

# Part II: Simulate data with missing passage counts under the "Weekends Off" scenario; fit a SN-NB arrival model #####
  # load libraries and set your directories (customize this to your machine)
    library(R2WinBUGS)
    library(sn)
```

```
  setwd("C:\\Users\\sureshsethi\\Desktop\\TestRuns") # WinBUGS models need to be stored here
  # set WinBUGS directory so computer knows where to find the program
  bd <- "C:\\Users\\sureshsethi\\Documents\\WINBUGS14"
# Simulate "true" data modeled after Pacific Salmon runs: Normal arrival curve model, run size = 10,000 fish,
# negative binomial process error
  set.seed(0) # set random seed if desired
  # Normal-curve shaped arrival model parameters
    true.day.v <- 1:365 # model time steps, modeled as 24 hour days, e.g. Julian days
    skew <- 0 # force skew to zero
    run.mu <- 185 # results in a mode on approximately 4th of July peak run
    run.sd <- 7.5 # produces a run of about 40 days long
    S <- 10000 # 10k fish run
  # simulate run arrival under the arrival model
  # for convenience the sn package is used, defining a skewnormal with shape = 0, i.e. Normal distribution
    true.run.v <- S*(psn(true.day.v,xi=run.mu,omega=run.sd,alpha=skew,engine="biv.nt.prob") -
        psn(true.day.v-1,xi=run.mu,omega=run.sd,alpha=skew, engine="biv.nt.prob"))
  # simulate observed data under process error
    real.dat.v <- rnbinom(n=length(true.run.v),mu=true.run.v,size=7) # overdispersion based on Pacific salmon data
      # note, ignore NA warnings here as these are due to rounding issues producing very small negative counts in tails
# Simulate observations at the weir with missing data following the "Weekends Off" scenario (observe 5, miss 2 days, ...)
  obs.start <- 160 # first day to observe at weir, e.g. June 9th by this example
  obs.end <- 210 # last day to observe at weir, e.g. July 29th by this example
  obs.dat.v <- data.frame(Date=true.day.v[obs.start:obs.end],Count=real.dat.v[obs.start:obs.end])
  # include known-zero passage days
  # Here, zero-passage dates before observation are equivalent to May 24 and 25, and post-observation
  # zero-passage dates are equivalent to August 13 and 14
    obs.dat.v <- rbind(data.frame(Date=144:145,Count=0),
      obs.dat.v,data.frame(Date=225:226,Count=0))
  # remove "weekend" data, i.e. sample for five days, take two off, sample five days...
    `%notin%`<- Negate(`%in%`)  # helper function
    knockout.date.v <- obs.dat.v[1:nrow(obs.dat.v) %in%
      c(seq(from=6,to=nrow(obs.dat.v),by=7),seq(from=7,to=nrow(obs.dat.v),by=7)),"Date"]
    knockout.passage.v <- obs.dat.v[obs.dat.v$Date %in% knockout.date.v,"Count"] # missed passages
    obs.dat.v <- obs.dat.v[obs.dat.v$Date %notin% knockout.date.v,]
# Take a look at the simulated "true" data and "observed" data with missing passage dates
  plot(x=obs.dat.v$Date,y=obs.dat.v$Count,xlim=c(125,265),ylim=c(0,1.5*max(obs.dat.v$Count)),type="p",col="red",
    pch=19,bty="l",xlab="Sample date (Julian day)",ylab="Passage count (fish)")
  points(x=knockout.date.v,y=knockout.passage.v,pch=19)
  lines(x=true.day.v[150:250],real.dat.v[150:250],type="l")
  lines(x=true.day.v[150:250],true.run.v[150:250],type="l",col="red",lwd=2)
  legend(x="topright",legend=c("Arrival model","Realized passage with process error","Missed passage date"),
    lty=c(1,1,1),lwd=2,seg.len=c(2,0,0),pch=19,pt.cex=c(0,1.25,1.25),col=c("red","red","black"),bty="n", cex=.8)
# Package data for WinBUGS and fit the skew-Normal arrival with Negative Binomial
# process error statistical arrival model
  # bundle data for WinBUGS
    win.data <- list(Date=obs.dat.v$Date,Count=obs.dat.v$Count,T=length(obs.dat.v$Count),invl=20)
    s.date <- win.data$Date[1]
    e.date <- win.data$Date[length(win.data$Date)]
  # add in the desired missing dates for prediction
    `%notin%`<- Negate(`%in%`)  # helper function
    miss.dates <- seq(from=s.date,to=e.date,by=1)[seq(from=s.date,to=e.date,by=1) %notin% win.data$Date]
```

```
  win.data$MDate <- miss.dates
  win.data$M <- length(miss.dates)
# specify 'inits' function for WinBUGS
  inits <- function(){list(
    xi_date_run=win.data$Date[win.data$Count==max(win.data$Count,na.rm=T)][1]+runif(1,-2,2),
    omega_date_run=12+runif(1,-1,1),
    alpha_date_run=0+runif(1,-2,2),
    theta=5+sample(c(-2,-1,1,2),1),
    logS=log(sum(win.data$Count))
    )}
  # specify parameters to track
   params <- c("xi_date_run","omega_date_run","alpha_date_run","theta","S","RunSize")
  # MCMC settings:
  nc = 2; ni = 10000; nb = 5000; nt = 10
  # Run Gibbs sampler to implement MCMC estimation of the model under Bayesian specification
  # Output true realized run size and estimated run size (i.e. sum of imputed passage + observed passage)
  fit <- bugs(data=win.data,inits=inits, parameters=params,
    model = "skewNormalArrival.NegativeBinomialErrorProcess.txt", n.thin=nt, n.chains=nc,
    n.burnin=nb, n.iter=ni, debug=F, bugs.directory=bd, digits=7, working.directory=getwd() )
        (results <- data.frame(TrueRunSize=sum(real.dat.v,na.rm=T),TotalObsevedPasage=sum(obs.dat.v$Count,na.rm=T),
  EstimatedRunSize=median(fit$sims.list$RunSize,na.rm=T)))
############################## End Part II ###########################################################################
```

Supplement 2. Sensitivity analysis for the inclusion of known zero passage dates

To facilitate fitting the statistical arrival models and prevent potential problems with numerical overflow resulting from extremely small predicted estimates at the tails of the run, we inserted zero passage dates before and after each simulated run curve. As discussed in the main text, this represents *a priori* knowledge that there are dates on either end of the run where zero passage can be assumed with a high degree of confidence.

To analyze the sensitivity of estimated run size to the placement and number of zero passage dates, four zero passage date placement schemes were applied to the first 20 datasets in each set of 150 simulated datasets for missing data scenario (see "*Performance testing*" in the Methods section of the main text). Simulated data sets were constructed such that the run (i.e. non-zero passage dates) occurred over Julian days 165-205. In Scheme I, implemented in analyses in the main text, zeros were placed on calendar days 144, 145, 225, and 226, bracketing the run data on either side by 20 days on average. Scheme II tripled the number of zeros to include calendar dates 140-145 and 225-230. Schemes III and IV follow I and II in structure, but move the zeros further from the center of the fish run. Scheme III places zeros on calendar dates 134, 135, 225, and 226 and scheme IV on 130-135 and 235-240. The average percent bias of the estimated total run size in each data scenario across statistical model and zero scheme is summarized in Table S1.1.

Of the four schemes examined here, none resulted in a consistently higher or lower average percent bias across data scenario or fitted statistical arrival model. As may be expected, scenarios with missing passage observation in the tails of runs were more sensitive to alternative zero passage date placements, demonstrated by the broader range in percent bias across zero schemes observed in these cases. For example, the greatest range in the percent bias across zero passage date placement schemes occurred when fitting a skew-Normal arrival – Normal process error statistical model to data simulated under the Normal arrival curve (5.9% range in percent bias) and skew-Normal arrival curve (5.91 % range in percent bias) with observation missing from the initial 15% of the run. In all other cases, the range in percent bias across zero schemes was substantially narrower. For example, the skew-Normal arrival – Normal process error model fit to the remaining data simulation scenarios demonstrated percent bias ranges $< 0.15\%$ across zero schemes (5 days on, 2 days off, Normal arrival: range = 0.03%; ± 2 days missing about the peak, Normal arrival: range = 0.07%; ± 2 days missing about the peak, skew-Normal arrival: range = 0.14%).

Furthermore, the choice of "best" statistical model for each data scenario as determined by smallest average absolute percent bias is consistent across zero schemes for four of the five data scenarios; in the remaining data scenario which indicated ambiguity between two statistical arrival models based upon the zero passage date placement scheme, the difference in average absolute percent bias between the best and second best performing statistical arrival models was less than 0.15% ("First 15% missed (Normal)" scenario, Normal arrival model – Normal process error and Normal arrival model – Negative Binomial process error statistical arrival models; Table S1.1). This indicates that selection of the best performing statistical arrival model using an expected bias criterion in the full analysis in the main text are robust to choice of zero passage date placement.

The results of this analysis suggest that statistical arrival model run size estimates and model selection inference are robust to the placement and number of zeros bracketing run data; however, we caution that asserted zero passage dates need be true zero passage dates and we encourage analysts to conduct sensitivity analyses specific to their datasets to verify robustness of estimates, particularly when faced with missing data in the tails of runs.

Table S1.1. Average run size percent bias (%) for each of the four zero schemes is shown for each data scenario[a].

Scenario (arrival model)	Average true run size	Fitted model	Average total run size estimate percent bias (%)			
			Scheme I	Scheme II	Scheme III	Scheme IV
Weekends off (Normal)	9,708	N-N	0.18	0.15	0.15	0.14
		N-NB	-0.19	-0.21	-0.19	-0.15
		SN-N	0.35	0.33	0.36	0.34
		SN-NB	**0.00**	**0.03**	**0.06**	**0.05**
First 15% missed (Normal)	10,107	N-N	0.27	**-0.17**	0.23	**-0.14**
		N-NB	**-0.13**	-0.19	**-0.12**	-0.26
		SN-N	5.24	1.33	7.22	4.82
		SN-NB	-2.45	-2.50	-0.41	-0.44
First 15% missed (skew Normal)	10,376	N-N	**0.57**	**0.18**	**0.55**	**0.22**
		N-NB	6.30	6.42	6.92	6.77
		SN-N	9.85	3.94	9.51	7.78
		SN-NB	0.61	0.88	3.23	1.40
±2 days about the peak missed (Normal)	9,930	N-N	0.87	0.92	0.87	0.94
		N-NB	**-0.03**	**-0.02**	**-0.04**	**-0.06**
		SN-N	0.76	0.83	0.79	0.80
		SN-NB	0.23	0.16	0.26	0.23
±2 days about the peak missed (skew Normal)	9,939	N-N	-1.12	-1.01	-1.12	-0.99
		N-NB	-8.24	-8.30	-8.29	-8.17
		SN-N	-0.75	-0.61	-0.66	-0.64
		SN-NB	**-0.26**	**-0.23**	**-0.28**	**-0.31**

[a]Bolded values indicate best performing statistical arrival models based on a criterion of lowest average absolute percent bias.

Table S1. Simulaton trial performance summary for estimation of underlying arrival model parameters[a].

Scenario (arrival model)	Fitted model	True location	Estimated location	True spread	Estimated spread	True dispersion	Estimated dispersion	True skew[b]	Estimated skew[b]
Weekends off (Normal)	N-N	185.2	185.2	7.59	7.55	7	.	0	.
	N-NB		185.2		7.62		7.1		.
	SN-N		186.1		10.66		.		1.88
	SN-NB		186.2		8.43		7.7		0.45
First 15% missed (Normal)	N-N	185.5	185.1	7.53	7.86	7	.	0	.
	N-NB		185.2		7.67		7.1		.
	SN-N		189.8		13.83		.		2.83
	SN-NB		186.1		9.22		7.6		0.77
First 15% missed (skew Normal)	N-N	185.1	184.8	7.56	5.57	7	.	2	.
	N-NB		184.1		5.74		6.1		.
	SN-N		188.9		10.30		.		3.55
	SN-NB		186.0		7.95		7.1		2.36
±2 days about the peak missed (Normal)	N-N	184.9	184.9	7.52	7.46	7	.	0	.
	N-NB		184.9		7.61		6.8		.
	SN-N		186.1		9.99		.		1.38
	SN-NB		186.3		8.39		7.2		0.47
±2 days about the peak missed (skew Normal)	N-N	184.5	185.0	7.51	5.20	7	.	2	.
	N-NB		185.1		5.45		3.2		.
	SN-N		185.8		7.61		.		2.77
	SN-NB		185.5		7.56		7.3		1.96

[a]Reported true and estimated location, spread, Negative Binomial dispersion parameter values are the mean parameter estimate across simulated datasets within a given scenario. Abbreviations: "N" = Normal, "NB" = skew-Normal, "NB" = Negative Binomial.
[b]True skewness was randomly drawn from the set $\{-2.0, 2.0\}$ during dataset simluations (see main text); as such, reported true and estimated skewness parameter values are the mean of the absolute value of parameter estimates across simulated datasets within a given scenario.