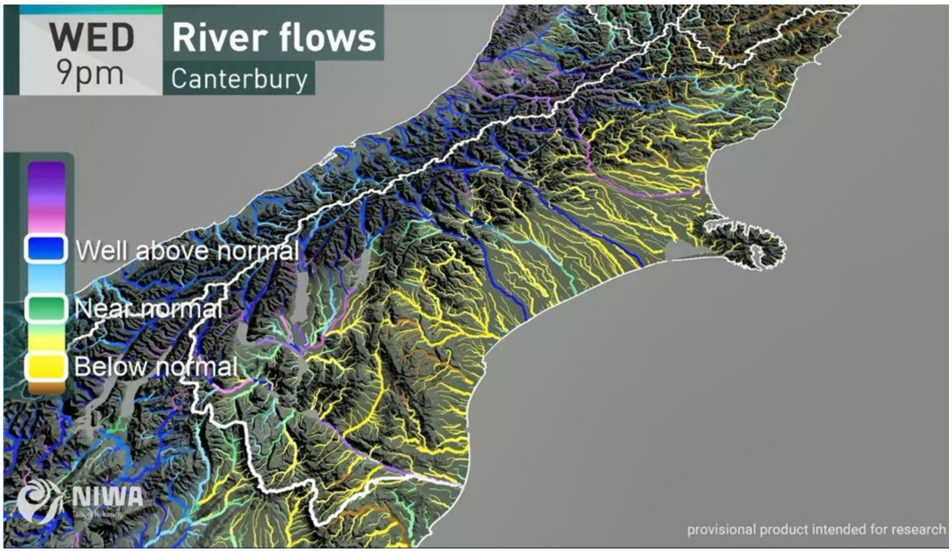
Preliminary evaluation of a national river flow forecast system for New Zealand

Prepared for NIWA SSIF

August 2020



Prepared by:
Jono Conway
Celine Cattoën
Kelsey Montgomery
Andrea Mari
Daniel Lagrava
Nava Fedaeff
Ude Shankar
Trevor Carey-Smith

For any information regarding this report please contact:

Jono Conway
Hydrological Forecasting Scientist
Hydrological Modelling
+64-3-440 0409
jono.conway@niwa.co.nz

National Institute of Water & Atmospheric Research Ltd
PO Box 8602
Riccarton
Christchurch 8011

Phone +64 3 348 8987

NIWA CLIENT REPORT No: 2020250CH
Report date: August 2020
NIWA Project: CARH2001

Quality Assurance Statement		
	Reviewed by:	Emily Lane
	Formatting checked by:	Rachel Wright
	Approved for release by:	Helen Rouse

Whilst NIWA has used all reasonable endeavours to ensure that the information contained in this document is accurate, NIWA does not give any express or implied warranty as to the completeness of the information contained herein, or that it will be suitable for any purpose(s) other than those specifically contemplated during the Project or agreed by NIWA and the Client.

Contents

Tables

Figures

Executive summary

The New Zealand (NZ) river flow forecasting system is a first attempt at producing and communicating flow forecasts for all rivers in New Zealand. The system has been running reliably for over a year now, generating categorical forecast flow information from convective-scale weather forecast output with a 48-hour lead time.

This report describes the design and components of the forecast system and lays out the framework used to evaluate the forecasts.

Validation of the forecasts is a work-in-progress, but initial results indicate that:

- Categorical flow forecasts (with six categories from well below normal to extremely high) can be made with reasonable skill at most sites, with mean absolute error around half a flow category.
- When assessed across all flow regimes (low, normal, high), errors in the hydrological model simulation and antecedent conditions dominate errors associated with forecasting precipitation for short lead times (up to 2 days). Therefore, increases in forecast accuracy are likely to come from improvements to the hydrological model and/or antecedent hydrological conditions used as forecast initial conditions (e.g., more accurate estimates of spatial and temporal patterns of actual precipitation and soil moisture).
- There is some indication that the model systematically underestimates high flows, but this should be assessed with a larger suite of sites and over a longer period that includes more flood events.
- Bias-correction of absolute flows using Flow Duration Curve (FDC) bias-correction substantially reduces the error and bias of most sites, and the extension of this method to sites without observed flow records should be pursued.

A detailed evaluation of forecast performance is being planned at a larger suite of gauged stations using a longer reforecast. This will enable reasons for good and poor performance to be properly diagnosed and the system to be improved. In parallel, engaging with potential users and stakeholders to acquire critical feedback and refine the system’s usefulness will continue to guide future development directions.

1 Background and aims

In New Zealand, flooding is the most frequent natural disaster while hydro-electricity is the main source of renewable energy. A national river flow forecasting system has potential benefits not only for hazard management and disaster risk reduction through early awareness of potential floods, but also for hydropower operation, recreation, irrigation planning and regulatory/monitoring activities.

The New Zealand river flow forecasting system[1] has been developed to produce and communicate categorical river flow forecasts at the national scale for ~60,000 river reaches (Strahler order 3 and above). The aim of the system is to complement existing regional flood forecasting systems and to provide New Zealand decision makers with a coherent overview of the current and future river flows.

In this report, we present an evaluation framework for the 48-hour categorical river flow forecasts along with a preliminary evaluation of forecast performance based on forecasts produced in the prototype operational system between October 2018 and August 2019. The performance is assessed at 43 sites that are available to NIWA in real time and meet basic suitability checks (Appendix A). The objectives of this study are to i) conduct a first performance assessment of the operational system prototype, ii) identify possible configuration issues impacting performance, and iii) understand the impact of operational data ingestion lag time to the forecasts. This work is the precursor of a future rigorous offline reforecast experiment which will assess performance at a larger number of flow sites (~400) and over a longer period (2.5 years of available forecast archive).

A specific evaluation of the hydrological model component is out of scope for this report as a larger sample of flow sites and longer forecast period is needed before conclusions can be drawn around hydrological process representation, particularly for forecasts of extreme high and low flow events. Such an evaluation will be undertaken in combination with a reforecast evaluation currently underway.

[1] https://niwa.co.nz/climate/research-projects/river-flow-forecasting

2 Methods

2.1 National-scale river flow forecast framework

National-scale river flow forecasts are produced by coupling the New Zealand Water Model's hydrology model, NZWaM-Hydro, to output from NIWA's high-resolution convection-permitting numerical weather prediction model, NZCSM (Figure 1). A 'best estimate' of the current hydrological conditions (i.e., river flow, soil moisture, shallow groundwater and snowpack storage) used as initial conditions for the forecasts is provided by model simulations driven by gridded observed climate data (Virtual Climate Station Network; VCSN (Tait et al. 2006)). Ensemble flow forecasts are produced every 6 hours, providing hourly hydrographs for more than 60,000 river reaches (Strahler order 3 and above[2]) for up to 48 hours lead time.

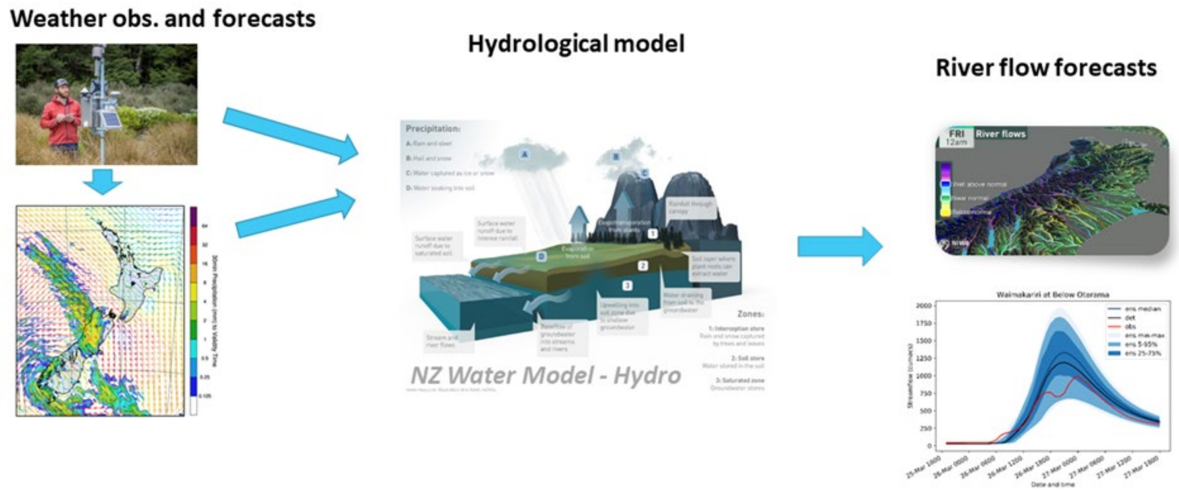


Figure 1: Components of the New Zealand river flow forecasting system.

2.1.1 Weather forecast model - NZCSM

NIWA's weather forecasts are generated by the New Zealand Convective Scale Model (NZCSM), a local implementation of the UK Met Office Unified Model System (UM). NZCSM is run as a deterministic model, with a grid resolution of 1.5 km and at a lead time of 48 hours. NZCSM takes its forcing from the New Zealand Limited Area Model (NZLAM), a regional weather model run at a 4.4 km resolution that uses lateral boundary conditions from the global version of the UM run by the UK Met Office. The 1.5 km grid resolution of the NZCSM allows an accurate representation of the New Zealand topography, which is especially beneficial in mountainous regions. NZCSM forecasts are issued four times a day, at 00:00, 06:00, 12:00 and 18:00 UTC.

[2] The Strahler order describes how large a river is based on the size of the branches of that are upstream of it. The smallest streams are order 1 and the largest rivers in New Zealand are order 8.

2.1.2 Observed weather - VCSN

The VCSN (Tait et al. 2006) interpolates observed meteorological values onto a grid covering New Zealand at a 5 km spatial resolution at a daily time step. A mass correction for hydrological modelling is necessary to overcome underestimation of precipitation in mountainous regions (Andréassian et al. 2010, Beck et al. 2019) The mass correction was calculated by comparing rainfall and long-term streamflow records, and correcting rainfall to ensure mass balance (Woods et al. 2006).

2.1.3 Hydrological model - NZWaM-Hydro

The New Zealand Water Model (NZWaM) is NIWA's framework for national river modelling; it encompasses model geospatial data (geofabric), hydrological models (quantity or quality), and applications (e.g., river flow forecasting, climate change scenarios).

NZWaM-Hydro is a distributed hydrological model based on TOPMODEL concepts of runoff generation controlled by sub-surface water storage. It combines a water balance model within each sub-catchment, with a kinematic wave routing algorithm (Beven et al., 1995, Goring, 1994). NZWaM-Hydro is set up to provide natural flow information for all of New Zealand and replicates the strong environmental diversity of New Zealand catchments (McMillan et al., 2016). The model simulates natural river flows and does not currently account for irrigation or dams.

In order to use a consistent approach in gauged and ungauged catchments, NZWaM-Hydro is uncalibrated to observed historic flows. Instead, model parameters are based on geospatial information (Hydro-Geofabric) drawn together by NIWA and other Crown Research Institutes, including Manaaki Whenua – Landcare Research and GNS Science about land use, soil and groundwater properties. Model parameters are therefore independent of biases in input (e.g., rainfall data) and are estimated using the same method in both gauged and ungauged catchments. The current version of operational systems uses digital river segments from the River Environments Classification v1 (RECv1[3]), retaining only stream segments of Strahler order 3 or larger. This results in over 60,000 river reaches being simulated across New Zealand.

2.1.4 Categorical flow conditions

To help remove hydrological model biases, simulated flow at each river reach is reported in six flow categories (e.g., well below normal, below normal) relative to hourly flow percentiles from a climatological simulation (Table 1).

Table 1: Percentiles of hourly flow data from climatology used to define flow categories.

Hourly Flow Percentile	Flow Category
> 99%th	Extremely high
90 - 99%th	Well above normal
66 - 90%th	Above normal
33 - 66%th	Normal
10 - 33%th	Below normal
0 - 10%th	Well below normal

[3] downloaded from: https://data.mfe.govt.nz/layer/51845-river-environment-classification-new-zealand-2010/

The modelled climatological percentiles were produced from a 40-year simulation with an identical model set-up to the forecast system using VCSN (Tait, et al., 2006) as climate input. To ensure the forecasted flows are consistent with the model climatology, the precipitation input from the NZCSM forecasts are bias corrected against the VCSN using a quantile matching procedure that accounts for a variable correction at different rain rates (Cattoën et al. 2016).

2.1.5 Ensemble hydrographs

To provide an estimate of the uncertainty associated with forecasts, an ensemble approach is taken (Cloke and Pappenberger 2009). In the ensemble approach, 50 different forecasts are made for each river reach, using small spatial and temporal variations in rainfall, soil moisture and baseflow to estimate uncertainty in forecast river flow (Clark and Slater 2006, Clark et al. 2008). The variations to each model state are randomly generated but are spatially and temporally correlated, attempting to mimic the variation observed in the real world. A single deterministic simulation is also made with no perturbations to rainfall, soil moisture and baseflow.

Ensemble flow information can be used to calculate the likelihood of flow conditions in each category. Figure 2a show the unperturbed deterministic forecast in black along with the different percentiles of the ensemble forecast in blue (e.g., median, 25th and 75th) and the actual flow that eventuated in red.

Both the deterministic and ensemble streamflow forecasts are processed into categories. Figure 2b shows that despite the large variability in absolute flow forecasts (Figure 2a), all ensemble members correctly forecast the “extremely high” category flows that occurred on the 27th March.

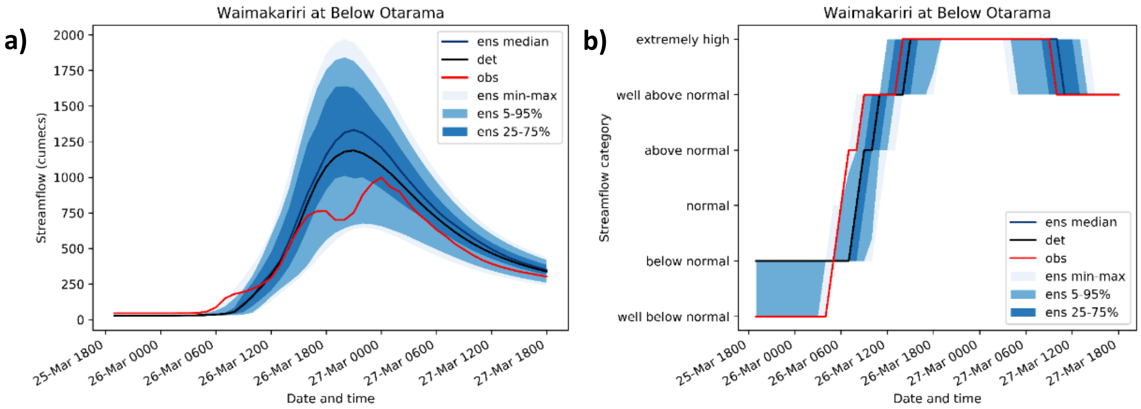


Figure 2: Raw forecast absolute flow (left) and categorical flow (right) along with observed flow that eventuated during a ~2.5yr-3yr return period flood event. Deterministic forecasts are shown in black and different percentiles of the ensemble forecasts are displayed in blue.

2.1.6 Visualisation of categorical river flow forecasts

To visualize and communicate flow forecast information, we produce a video showing the next 48 hours forecast daily using the Presentation Cartography software[4]. We created two forecast views: the basin overview (Figure 3a), and river network view (Figure 3b). The basin overview covers the whole country and colours each sub-catchment according to the flow in the corresponding river reach. The river network view shows each region in turn using three levels of river thickness to differentiate between river sizes of Strahler order 3, 4 and 5 to 7. Forecast flow videos are

[4] https://presentationcartography.com/

automatically rendered daily within the computational framework presented above. The presentation of categorical flow forecast information through videos is being tested by a group of stakeholders. Based on feedback from the stakeholders, an overview of accumulated rainfall has been added to the start of the videos.

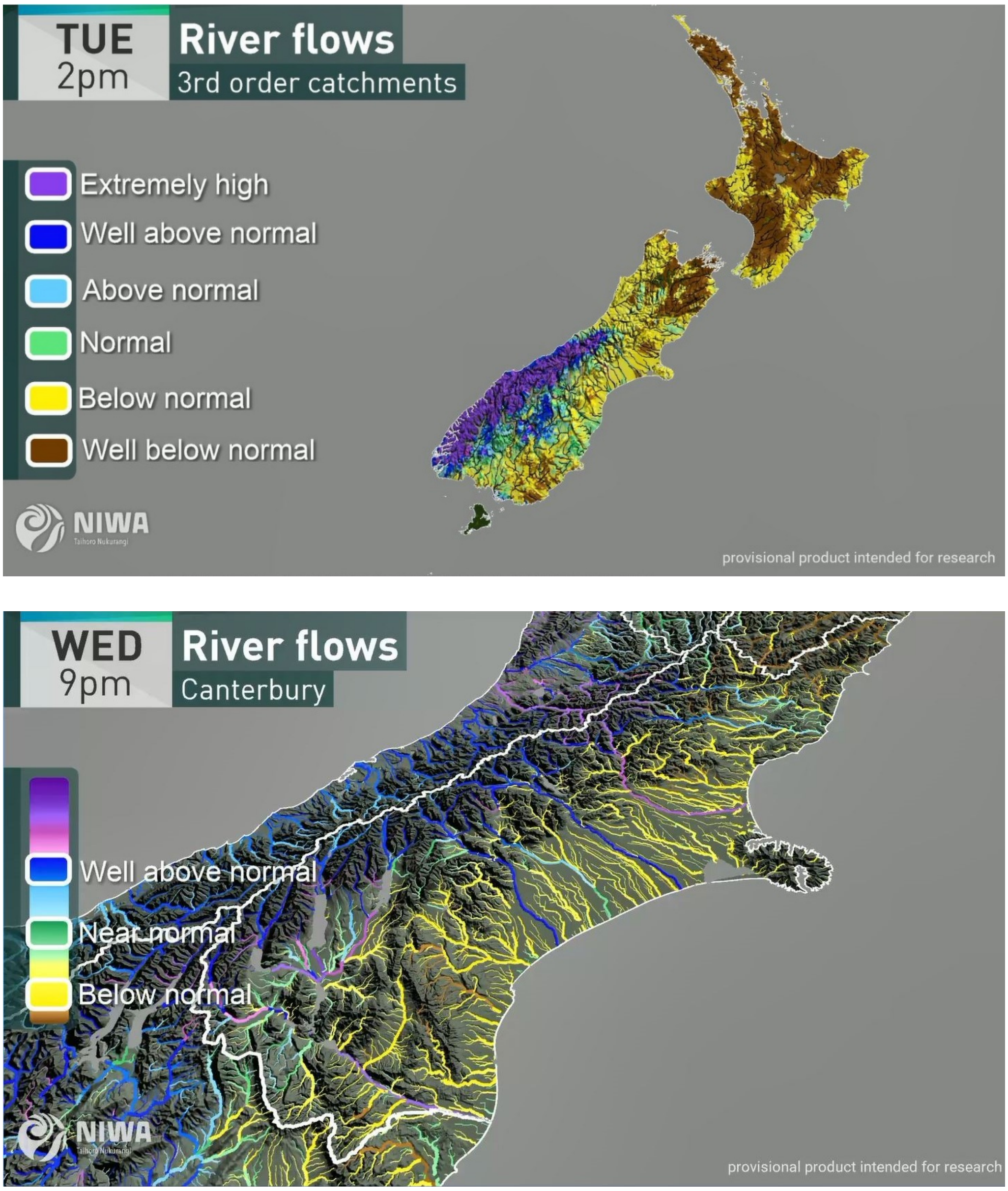


Figure 3: Snapshots of video forecasts displaying categorical flow conditions for sub-catchments (top) and river reaches (bottom) during March 2019 flood event.

2.1.7 Bridging the gap between observed data and forecast

The initial hydrological conditions needed for the Forecast simulations (flow, soil moisture, baseflow, etc.) are generated by a model simulation forced by gridded daily weather observations (VCSN) that represents a best estimate of the current hydrological conditions (Figure 4). However, the operational VCSN cannot be generated until one or more days after the present forecast period due to lags in the availability of some weather station data. Because of this lag, a 'Bridging' simulation is made to take the initial conditions from the VCSN suite and bring them up to the current forecast time. This Bridging suite uses rainfall and other weather output from previous NZCSM forecasts. The VCSN and Bridging simulations are updated once a day, with the 1800 UTC (6pm NZST) Forecast simulation taking its initial conditions from the most recent Bridging simulation. The 0000, 0600 and 1200 UTC forecasts take their initial conditions from the previous forecast simulation. As noted in Section 2.1.4 the NZCSM precipitation used in the Bridging and Forecast simulations is bias-corrected to resemble VCSN precipitation in order to keep the Bridging and Forecast simulations consistent with the VCSN simulations.

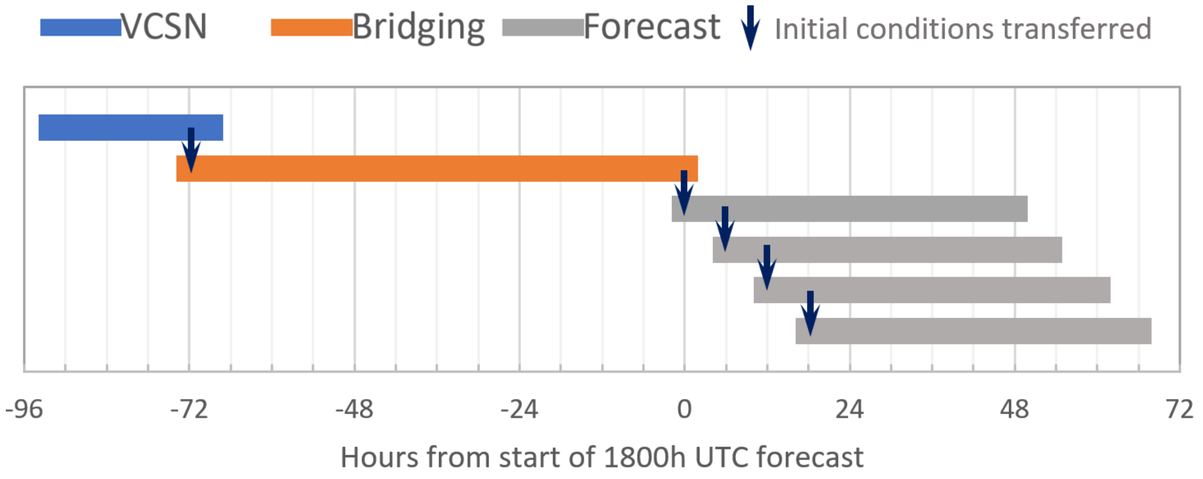


Figure 4: Schematic showing flow simulations used to generate initial conditions for forecasts.

2.1.8 Forecast Flow Duration Curve (FDC) bias-correction

We explore the potential of an alternative method to remove hydrological model biases and produce absolute discharge forecasts instead of categorical forecasts (e.g., 60 m^3/s instead of ‘above normal’). A flow duration curve (FDC) bias correction procedure rescales flow forecasts to improve predictions of the streamflow values at gauged sites (Farmer et al. 2018). The FDC bias correction procedure has two steps. Firstly, observed and simulated FDC are created using an observed flow timeseries and a 40-year simulated flow climatology, respectively. In the second step, raw simulated values from a forecast are mapped to a flow percentile (exceedance probability) on the simulated FDC. The absolute flow value (in m^3/s) corresponding to the same flow percentile in the observed FDC is then used to create the bias-corrected values for each timestep (Figure 5). This second step is repeated for each timestep in a forecast using the same simulated and observed FDC.

Bias-correcting daily hydrological simulations with FDC estimates has shown the greatest improvements for the mid-range flows in New Zealand (McMillan et al. 2016). In this report, we investigate the impact on the forecast performance of applying an FDC bias-correction procedure to hourly flows at gauged sites only, because this technique relies on having an observed FDC. The shape of an FDC is characteristic for a given basin and is controlled by precipitation and basin characteristics. Machine learning methods could potentially be used to estimate FDC in ungauged rivers (Booker 2012, Worland et al. 2019), providing a method to reliably estimate absolute flows (in m^3/s) for all locations in New Zealand.

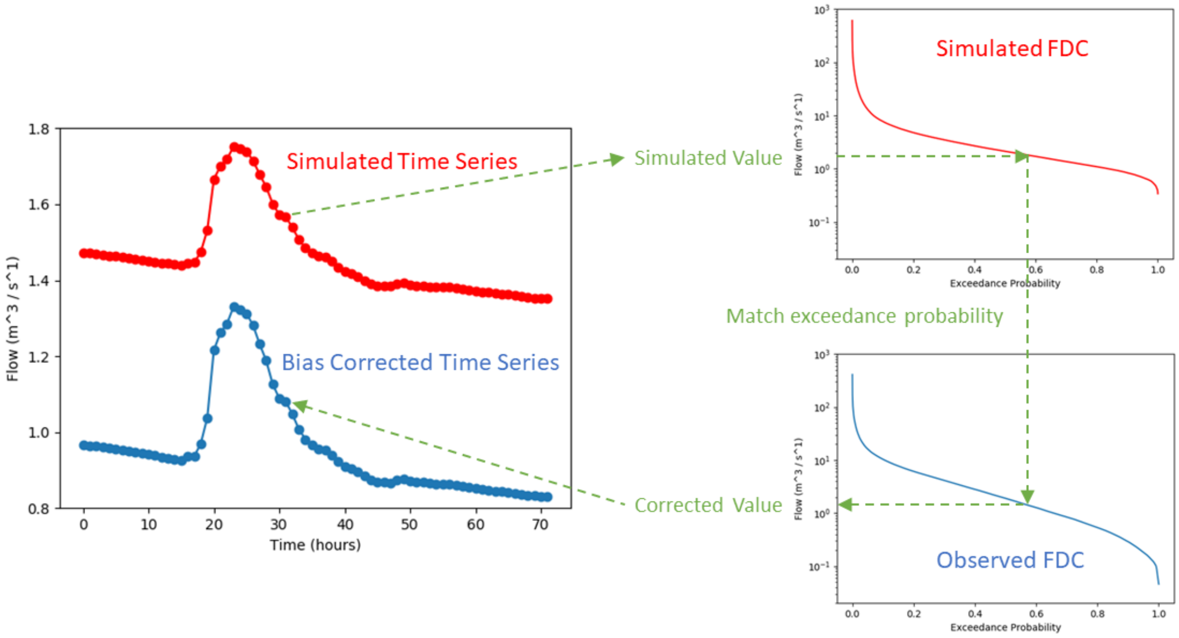


Figure 5: Bias correction of raw simulated forecast flow time series using a simulated and observed FDC.

2.2 Evaluation framework

The archive of prototype forecasts began in October 2018 and forecasts up till August 2019 are evaluated here at 43 sites where data is available to NIWA in near-real time. Note that these sites include ones with managed river flow and/or water abstractions, both of which will degrade model performance. Sites located on very small headwater streams that are not represented in the Strahler 3 digital river network are not included in the 43 sites. Note that ensemble results for the West Coast region are not available due to a software bug in the ensemble generation within the hydrological model that was identified during this study (which has now been resolved in the operational pipeline). In that respect, we fulfilled the second objective of this work which was to identify possible configuration and model issues impacting forecast performance. Ensemble forecast results for the West Coast region will be analysed in future reforecast experiments.

Simulated flow is compared to two different baselines:

- **Observed** streamflow. Forecast errors against observed streamflow combine errors from forecast rainfall and meteorology, as well as errors in the hydrological model and initial hydrological conditions (e.g., antecedent river flow, soil moisture, storage in shallow groundwater and snowpack).
- **Pseudo-observed** streamflow, which is the flow simulated by the VCSN simulation (Figure 4). Forecast errors against pseudo-observed streamflow are dominated by the errors in forecast precipitation as the same hydrological model is used for both simulations and the pseudo-observed stream flow is used as the initial conditions for the forecast.

For evaluation, timeseries of categorical observed flow are calculated for each site in the same way as for simulations, using the observed flow series to generate flow percentiles. Note that the observed record at each site does not always cover the period of 40-year climatology, and this may cause issues with categorical biases (see Section 3.3).

2.2.1 Forecast accuracy and bias

Basic performance measures to assess forecast accuracy and biases include:

- Mean absolute error (MAE) for deterministic forecasts and its counterpart the ranked probability score (RPS)[5] for ensemble forecasts. The RPS measures how well the ensemble forecast predicted the category that the observation fell into and a perfect score is 0. For an ensemble size of 1, the RPS is equal to the MAE.
- Mean bias, which is calculated as the mean of the simulated minus observed (or pseudo-observed) flow. For ensemble forecasts, the bias is calculated for the mean of the ensemble.

Forecast bias and error are compared to catchment characteristics including:

- stream order from the RECv1 digital stream network.
- upstream catchment area from the RECv1 digital stream network.
- mean flow from observed flow records at given site.
- specific discharge, which is a measure of catchment wetness and is calculated as mean flow / catchment area.

2.2.2 Threshold exceedance scores

As a first step towards evaluating the system performance in flood conditions, binary threshold exceedance scores are evaluated for flows exceeding the 80th percentile climatological flow. The 80th percentile is chosen as a threshold in order capture enough events in the short record we have in this study. Threshold exceedance metrics calculated include:

- Event frequency = Here we split each forecast period into 6-hour windows and count an event if flow exceeds the threshold at any point in the window. Events are counted separately for observed and forecasted flow over the same time windows.
  - A perfect forecast will have the same frequency of observed and forecasted events.
- Frequency bias = number of forecasted events / number of observed events
  - Frequency bias less than 1 indicates that high flows are forecast less often than they occur, while frequency bias greater than 1 indicates high flows are forecast more often than they occur.
- Hit rate = number of correctly simulated exceedances / number of actual exceedances.
  - Perfect == 1
- False alarm rate = number of false alarms / number of actual non-exceedances.
  - Perfect == 0

A more thorough evaluation of system performance in flood conditions will be undertaken using output from a reforecast experiment that is underway. The reforecast experiment will generate flow

[5] https://www.cawcr.gov.au/projects/verification/

simulations using the archive of NZCSM weather forecasts for the period June 2017 to April 2020. The reforecast will be assessed against a much larger number of sites, including data from regional councils that is not currently transferred to the NIWA archives operationally.

The longer reforecast period will provide a much larger number of flood events required to evaluate high-flow exceedance performance in a more statistically meaningful way. This includes evaluating the reliability of ensemble forecasts in high-flow conditions, a key aspect of forecast performance that is not addressed in this report. The ensemble reliability describes whether the probability of a forecast high-flow event (calculated from the ensemble) reflects how often a high-flow event occurs in the real world. In a reliable forecasting system, during the times that most ensemble members forecast a flood (high forecast probability), high flow should occur most of the time. During the times that few ensemble members forecast a high-flow event (low forecast probability), high flow should only be observed some of the time.

3 Preliminary results

3.1 Overall performance compared to observed flow categories

In general, the deterministic forecast at most sites has a reasonable categorical error of between 0.4 and 0.6 flow categories, while a few sites have much larger errors (Figure 6). It is likely some of the poor performing sites are those with managed flow. Poor performance of the model in the eastern Bay of Plenty (high error and negative bias) was also found for the 40-year climatology simulations and is likely related to geology, which does not conform to some of the hydrological model assumptions.

There is a tendency for forecasts to have lower categorical flow than observed (blue tones in Figure 6), while sites in the north-west of the South Island have forecast flow higher than observed. The VCSN precipitation (used to bias-correct the NZCSM rainfall forecasts) in the Tasman region is known to be poor for large precipitation events, so the large errors in the region may be related to the precipitation input.

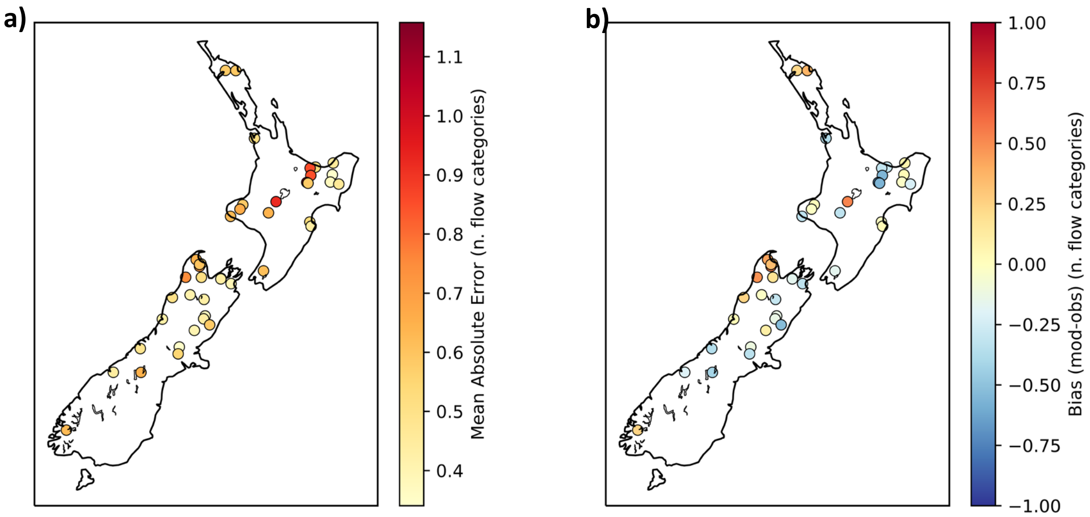

a)
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Mean Absolute Error (n. flow categories)
b)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
Bias (mod-obs) (n. flow categories)

Figure 6: Mean absolute error (a) and bias (b) of deterministic forecasts (at all hourly lead times) at selected sites. The units are number of flow categories (e.g., normal to above normal = 1 category).

3.2 Variation of performance with catchment characteristics

The variation of forecast error and bias is assessed with respect to catchment characteristics (Figure 7) to evaluate the suitability of the forecast framework for use in the diverse catchments within New Zealand. The limited number of sites, particularly those with high mean flow and high specific discharge (wetter) limit the conclusions that can be drawn, but, in general, no clear relationship between model forecast performance and stream order or catchment area can be discerned. Higher mean flow or specific discharge sites appear to have lower errors than drier sites. The poorer performance in drier catchments may come from the representation of hydrological initial conditions (e.g., soil moisture), precipitation or hydrological model errors. The planned reforecast exercise will assess these relationships at a much large number of sites (~400), allowing the robustness of and reasons for this trend to be fully assessed.

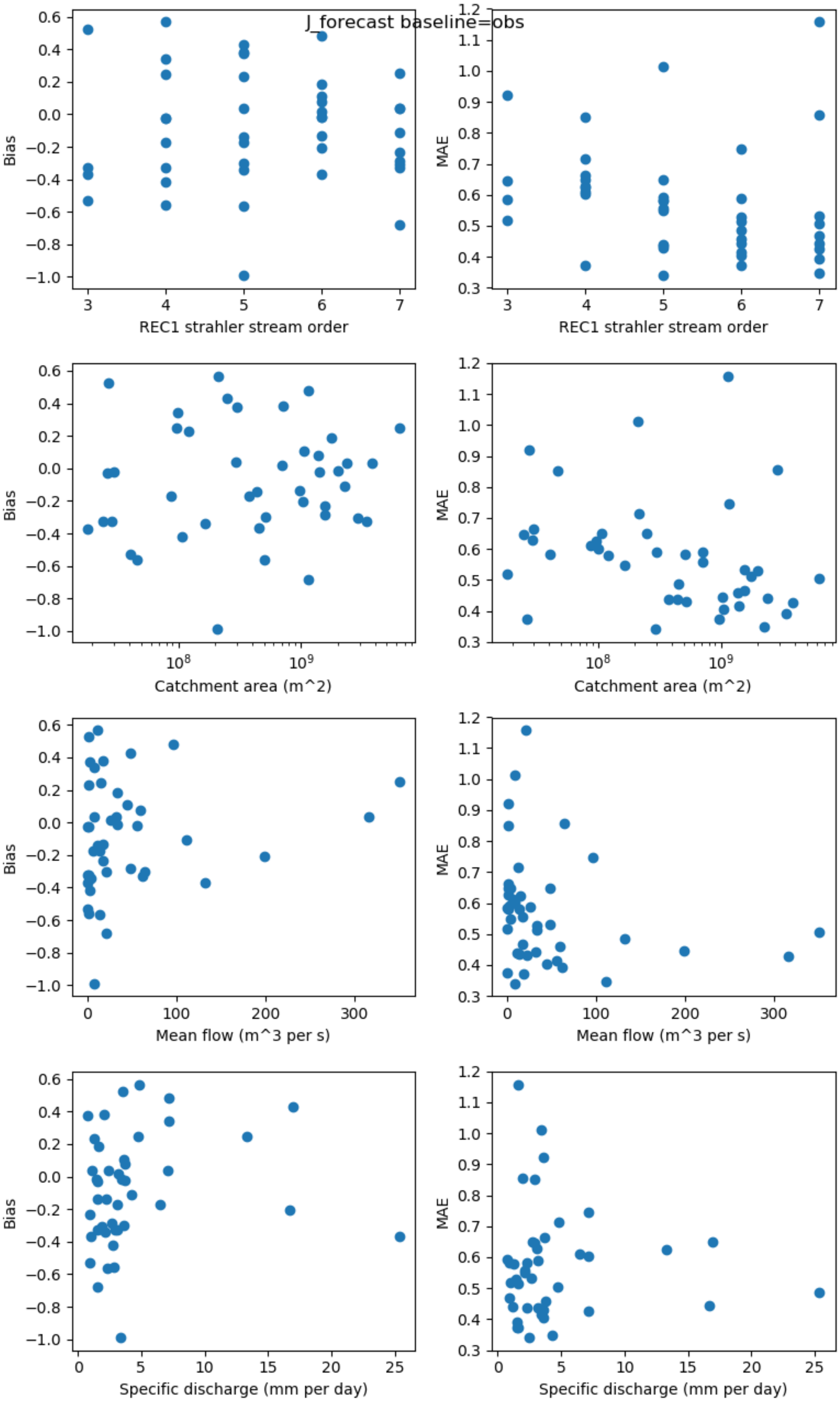


Figure 7: Bias (left) and mean absolute error (right) of deterministic forecasts (at all hourly lead times) against observed categorical flow at selected sites compared with catchment characteristics.

3.3 Performance across lead times compared to different baselines

We assess the forecast performance across lead times against two baselines i) observed flows and ii) pseudo-observed flow (the VSCN simulation where flow is forced by observed climate data).

The first approach i) includes errors from the meteorological model, hydrological model and initial hydrological conditions. The second approach ii) isolates the loss of skill due to errors in initial hydrological conditions and hydrological model, and focus on meteorological model errors and biases.

In i) the comparison to simulated and observed categorical flow (Figure 8a,c) shows:

- Ensemble forecast error is smaller than deterministic error (Figure 8a). This confirms our expectation that the ensemble approach is more robust than the deterministic.
- Error does not increase significantly with lead time at this short range (less than 48 hours) (Figure 8a). This indicates that errors in the hydrological model and initial condition uncertainty likely dominate forecast rainfall errors over the evaluation period. Smaller error for Forecast simulations compared to VCSN simulations are likely due to the smaller biases observed in the Forecast simulations (see next point).
- Most sites tend to underestimate relative flow against relative flow observations (Figure 8c), especially for the VCSN simulations. There are several possible reasons for this trend that will need to be investigated further using the reforecast experiment. It may be that as-yet undiagnosed biases remain in the hydrological model set up or there may be issues with the observed relative flow used for comparison (e.g., different length of record from climatological simulation used to create relative flow percentiles).
- Average flow tends to increase with lead time (Figure 8c,d). This indicates that NZCSM driven simulations are, in general, wetter than those driven with VCSN. There are various small differences in model set up that may contribute to this: different spatial resolution (5km vs 1.5km), the hourly disaggregation of rainfall (stochastic vs model simulated) or the estimation of solar radiation and/or ET. Further analysis will be undertaken to assess this feature.

In ii) the comparison of simulated and pseudo-observed flow (Figure 8b,d) shows:

- Ensemble flow errors grow slowly with lead time (Figure 8b), indicating it is likely we can forecast further into the future without major degradation of hydrological skill.
- Forecast error is less than half of total model error at a 48-hour lead time (compare Figure 8a and 8b). This emphasises that errors in the hydrological model simulation (including antecedent conditions and actual patterns of precipitation) dominate errors associated with forecasting precipitation for short lead times.

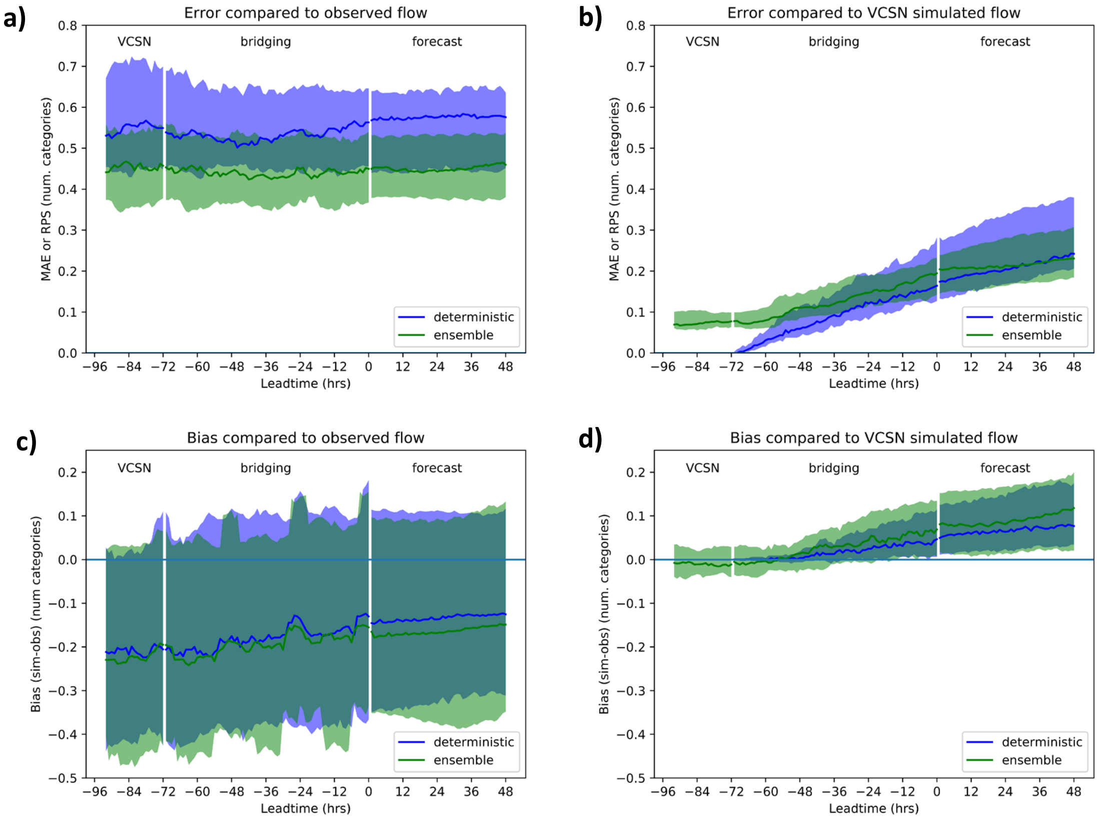


Figure 8: Categorical flow error (top) and bias (bottom) for VCSN, Bridging and forecast simulations compared to observed (left) & pseudo-observed (right) flow. Solid lines show median and shading the interquartile range of 37 sites excluding the West Coast region.

Note the Forecast suite at lead time 0h has higher bias and error than the Bridging suite at lead time 0h (Figure 8d). This is because the initial conditions from the Bridging suite are transferred once a day for the at 1800 UTC forecasts, so the other forecast cycles have a longer period of simulation with NZCSM input after the VCSN simulation. Isolating the performance of 1800 UTC forecasts shows that these simulations have lower error than the other forecast cycle times (Figure 9) and that the errors for the 1800 UTC forecast are consistent with the Bridging suite, confirming that there is no discontinuity between the Bridging and Forecast suite.

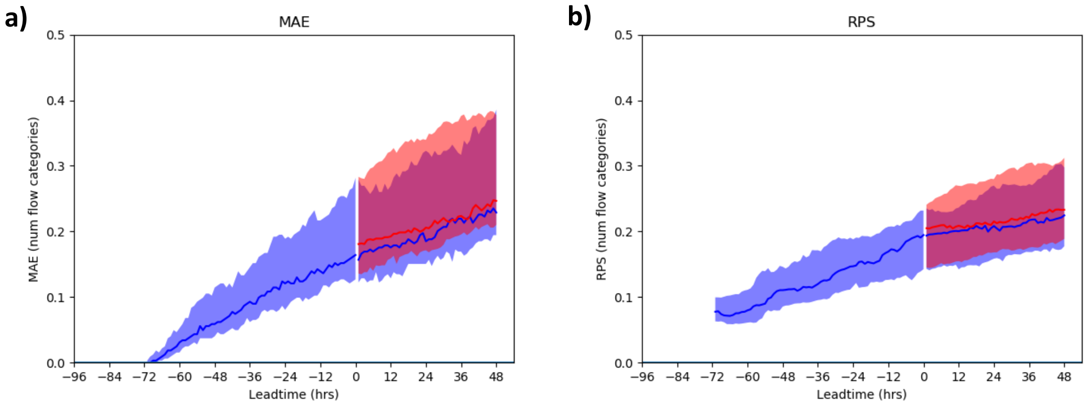


Figure 9: Categorical flow error compared to pseudo-observed flow. for Bridging and 1800 UTC Forecast (purple/blue) and other Forecast cycle times (red) for deterministic (a) and ensemble (b) simulations.

3.4 High-flow threshold exceedance scores

An initial evaluation of the performance of the forecasts for high-flow threshold exceedance against observed data (Figure 10) shows:

- The frequency of high flows (> 0.8) is generally under-forecast (frequency bias < 1) at most sites when using the deterministic simulation or the mean of the ensemble simulation (Figure 10a, b).

- Similarly, the hit rate is low for many sites (Figure 10b). Figure 11 shows the two metrics are closely related – at sites where more high flows are forecast (frequency bias > 1), the more likely it is that actual observed high flow events are forecast (hit rate closer to 1).

- Ensemble probabilities can be optimised to trade-off between hit rate, false alarms and frequency and produce more reliable forecasts. For example, when requiring less ensemble members to exceed threshold before event is defined (e.g., 25% probability) hit rate can be increased (median of 0.5 to 0.65) and event frequency improved (closer to observed) with only small increase in false alarms (Figure 10d)

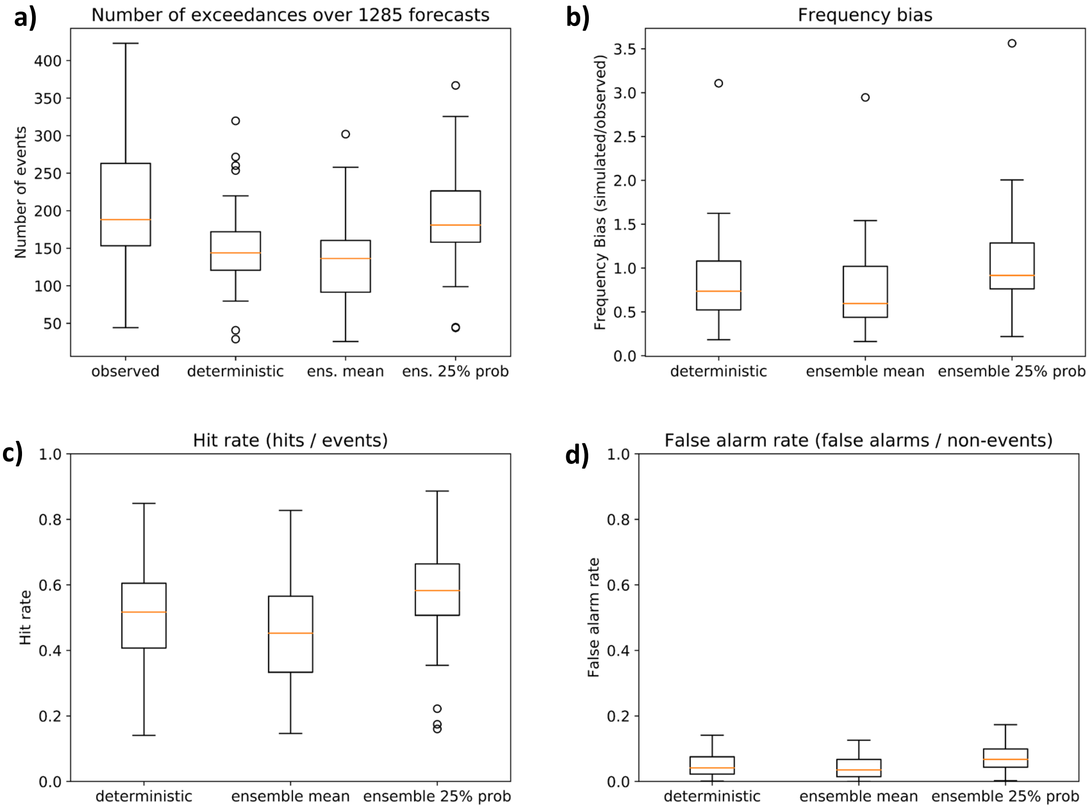


Figure 10: Event frequency (a), frequency bias (b), hit rate (c), false alarm rate (d) for threshold exceedance of 80th percentile relative flow. Boxplots show overall performance of 37 sites using deterministic forecast or when using ensemble mean or 25% probability to calculate exceedance. A 6-hour window is used to calculate exceedance in both observed and simulated flow.

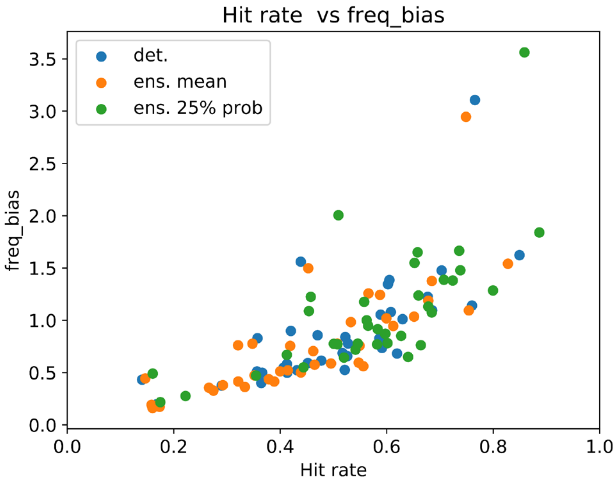


Figure 11: Scatter plot of threshold-exceedance scores (Hit rate and frequency-bias) for 37 sites using same methods as Figure 8.

3.5 FDC bias-corrected absolute forecasts

In order to make the forecast system more useful to stakeholders, forecasts of absolute flow using FDC bias-correction are being investigated. Currently FDC bias corrected absolute flows can only be provided at locations with sufficiently long observed flow records (a few hundred sites nationwide). Here we make an initial evaluation of whether FDC bias correction improves the simulation of absolute flows. Figure 12 shows that, overall, the error and bias of absolute forecasts are improved (smaller errors and bias closer to 0) with FDC bias correction. For most sites the error decreases, but the error is slightly increased for some sites that do not exhibit large biases prior to bias-correction. The improvement of absolute forecasts using FDC bias-correction further confirms our choice to present forecasts in categorical flow units relative to the flow climatology (similar to the FDC bias correction) in the current system. It also highlights the potential of future developments that would enable FDC bias-corrected absolute flow forecasts to be provided at all sites.

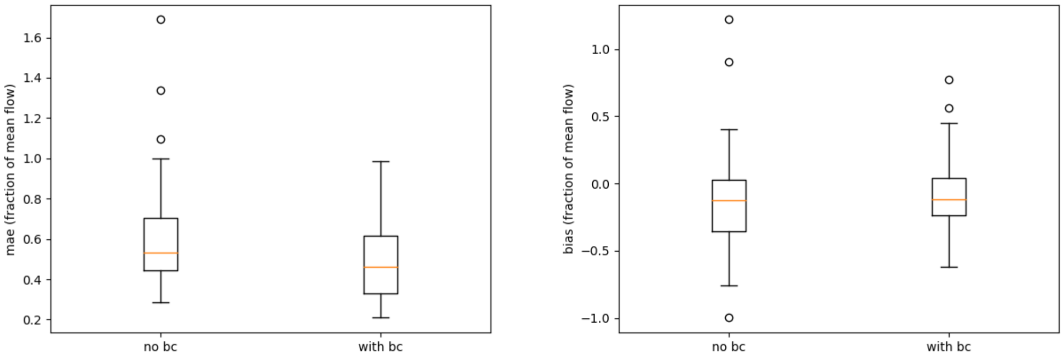


Figure 12: Mean absolute error (left) and bias (right) of raw deterministic forecasts ('no bc') and FDC bias-corrected forecasts ('with bc'). against absolute river flows at 43 selected sites for all hourly lead times. Note that while the errors are calculated in flow units (m^3/s) for each site, the plots show the results normalised using the mean flow at each site to enable comparison between sites.

4 Current and future developments

Several initiatives are being investigated to improve the flow forecasting system:

- Extending the validation to sites operated by regional councils to understand where and why the model performs well or not. This will guide future improvements.
- Verifying performance for flood flows using longer forecast archives (at least several years of forecasts). This is likely to come from a reforecast, which feeds weather forecasts from the last few years into the flow forecasting system to create a longer archive of forecasts.
- Decreasing the time lag before observed rainfall can be used.
- Including regional council rainfall data in the operational VCSN input. This is likely to substantially improve the simulation of flows in areas where VCSN is known to have issues with precipitation.
- Increasing the reliability of the ensemble forecasts (i.e., so that the ensemble always includes the observed flow and high flow events are not forecast too often). This includes investigating using:
  - statistical ensembles of bias-corrected rainfall (Cattoën et al. 2020)
  - ensemble forecasts from numerical weather prediction models (e.g., the 120-hour 4 km ensemble [NZENS] being tested)
  - lagged rainfall forecasts from successive forecast cycles
  - methods to include hydrological model uncertainty (e.g., different parameter sets).
- Optimising probability thresholds used to forecast high flow exceedance from ensemble flow information.
- Developing methods to reliably convert categorical flow to bias-corrected absolute flow in locations without observed data.
- Increasing forecast lead times using rainfall from several forecast models.

5 Summary

The NZ river flow forecasting system is a first attempt at producing and communicating national flow forecasts driven by a convective scale weather model. The system has been running reliably for over a year now, generating forecast flow information for all streams and rivers in New Zealand with 48-hour lead time.

Validation of the forecasts is a work-in-progress, but initial results indicate that:

- Categorical flow forecasts (in 6 categories) can be made with reasonable skill (errors around half a flow category) at most sites.
- When assessed across all flow regimes, errors in the hydrological model and antecedent conditions dominate errors in forecast precipitation for short lead times (up to 2 days). Therefore, increases in forecast accuracy are likely to come from improvements to the hydrological model and/or antecedent hydrological conditions used as forecast initial conditions (e.g., more accurate estimates of spatial and temporal patterns of actual precipitation and soil moisture).
- The forecast performance for high flows needs further investigation in a reforecast, with some indication the model systematically underestimates high flows.
- Improved forecast ensembles are needed before forecasts of exceedance of high-flow thresholds can be reliably provided.
- Bias-correction of absolute flows using FDC bias-correction substantially reduces the error and bias of most sites, and the extension of this method to sites without observed flow records should be pursued.

As the operational archive grows and quality-assured observed flow data becomes available, evaluation of forecast performance will be undertaken at a larger suite of gauged stations, enabling reasons for good and poor performance to be properly diagnosed. This evaluation will be supported by a dedicated reforecast experiment to produce historical flow forecasts using archived weather forecasts. In parallel, engaging with potential users and stakeholders to acquire critical feedback and refine the system’s usefulness will continue to guide future development directions.

Further information can be found at the Project website: https://niwa.co.nz/climate/research-projects/river-flow-forecasting

6 Acknowledgements

This project is funded through the NIWA MBIE SSIF Hazards Programme, Project CARH**01, with contributions from CARH**02 and CARH**03. Input from many regional councils and other stakeholders has been invaluable in the ongoing development of the system.

7 Glossary of abbreviations and terms

False alarm rate	Number of false alarms / number of actual non-exceedances.
Frequency bias	Number of forecasted events / number of observed events
Hit rate	Number of correctly simulated exceedances / number of actual exceedances.
Pseudo-observed	Flow simulated using observed weather information

8 References

Andréassian, V., Perrin, C., Parent, E. & Bárdossy, A. 2010. The Court of Miracles of Hydrology: can failure stories contribute to hydrological science? *Hydrological Sciences Journal,* 55**,** 849-856.

Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G. and coauthors 2019. MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bulletin of the American Meteorological Society,* 100**,** 473-500.

Booker, D. J. 2012. Comparing methods for estimating flow duration curves at ungauged sites. *Journal of hydrology,* v. 434-435**,** pp. 78-94-2012 v.434-435.

Cattoën, C., Carey-Smith, T. & McMillan, H. 2016. Operational flood forecasting with a bias corrected high resolution weather forecasts. *56th NZ Hydrological society, 37th Australian hydrology and water resources symposium, 7th IPENZ rivers group.* Queenstown, New Zealand.

Cattoën, C., Robertson, D. E., Bennett, J. C., Wang, Q. J. & Carey-Smith, T. K. 2020. Calibrating Hourly Precipitation Forecasts with Daily Observations. *Journal of Hydrometeorology,* 21**,** 1655-1673.

Clark, M. & Slater, A. 2006. Probablistic quantitative precipitation estimation in complex terrain. *Journal of Hydrometeorology,* 7**,** 3-21.

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P. and coauthors 2008. Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources,* 31**,** 1309-1324.

Cloke, H. L. & Pappenberger, F. 2009. Ensemble flood forecasting: A review. *Journal of Hydrology,* 375**,** 613-626.

Farmer, W. H., Over, T. M. & Kiang, J. E. 2018. Bias correction of simulated historical daily streamflow at ungauged locations by using independently estimated flow duration curves. *Hydrol. Earth Syst. Sci.,* 22**,** 5741-5758.

McMillan, H. K., Booker, D. J. & Cattoën, C. 2016. Validation of a national hydrological model. *Journal of Hydrology,* 541**,** 800-815.

Tait, A., Henderson, R., Turner, R. & Zheng, X. 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. *International Journal of Climatology,* 26**,** 2097-2115.

Woods, R., Hendrikx, J., Henderson, R. D. & Tait, A. 2006. Estimating mean flow of New Zealand rivers. *Journal of Hydology (NZ),* 45**,** 95-110.

Worland, S. C., Steinschneider, S., Asquith, W., Knight, R. & Wieczorek, M. 2019. Prediction and Inference of Flow Duration Curves Using Multioutput Neural Networks. *Water Resources Research,* 55**,** 6850-6868.

Appendix A List of sites used in evaluation

Table A-1: Details of sites used in evaluation.

Site name	Station ID (aka Tideda number)	River Environments Classification (REC) version 1 rchid	REC1 Strahler order	Catchment area (m^2)
Waipapa at Forest Ranger	47804	1007423	5	1.21E+08
Waitangi at Wakelins	3722	1007625	5	3.01E+08
Waitangi at SH Bridge	43602	2009679	3	1.82E+07
Motu at Houpoto	16501	4005116	6	1.38E+09
Whakatane at Whakatane	15514	4008284	7	1.56E+09
Rangitaiki at Te Teko	15412	4009262	7	2.89E+09
Motu at Waitangirua	16502	4016669	5	2.94E+08
Waihua at Gorge	15453	4017064	4	4.66E+07
Rangitaiki at Murupara	15408	4022486	7	1.15E+09
Whirinaki at Galatea	15410	4022892	5	5.07E+08
Waikohu at No 1 Br	19734	5009160	4	2.65E+07
Waipaoa at KanakanaiaC/W	19716	5010343	7	1.57E+09
Waitara at Tarata	39501	6004432	6	7.04E+08
Manganui at SH3	39508	6006851	4	2.98E+07
Punehu at Pihama	36001	6010817	4	2.89E+07
Whanganui at Te Porere	33347	7008385	3	2.72E+07
Makotuku at SH49A Br	33117	7015868	3	2.46E+07
Ngaruroro at Chesterhope Br	23150	8024658	6	2.00E+09
Tukituki at Red Br	23201	8026822	7	2.38E+09
Hutt at Kaitoke	29808	9008427	4	8.69E+07
Aorere at Devils Boots	52003	10001004	5	2.48E+08
Takaka at Kotinga Br	52901	10002960	5	7.13E+08
Waingaro at Hanging Rock	52904	10003224	4	2.13E+08
Anatoki at Happy Sams	52903	10004171	4	9.98E+07
Motueka at Woodstock	57009	10010534	6	1.76E+09
Rai at Rai Falls	58903	11011511	5	2.09E+08
Pelorus at Bryants	58902	11012028	5	3.78E+08
Wairau at Barnetts Bank	60109	11016543	7	3.43E+09
Wairau at Dip Flat	60114	11029639	5	5.18E+08
Karamea at Gorge*	95102	12003174	6	1.16E+09
Buller at Longford*	93202	12010223	6	1.40E+09
Buller at Te Kuha*	93203	12012479	7	6.35E+09
Grey at Dobson*	91401	12028095	7	3.83E+09
Whataroa at SHB*	89301	12042448	6	4.51E+08
Haast at Roaring Billy*	86802	12052272	6	1.03E+09
Acheron at Clarence	62103	13008668	6	9.73E+08
Waiau Toa/Clarence at Jollies	62105	13010715	5	4.40E+08

Site name	Station ID (aka Tideda number)	River Environments Classification (REC) version 1 rchid	REC1 Strahler order	Catchment area (m^2)
Stanton at Cheddar Valley	64610	13015628	3	4.06E+07
Hurunui at Mandamus	65104	13020391	6	1.06E+09
Waimakariri at Below Otarama	66403	13036997	7	2.25E+09
Selwyn at Whitecliffs	68001	13043804	5	1.65E+08
Forks at Balmoral	71129	13506748	4	1.07E+08
Spey at West Arm	79740	15027181	4	9.59E+07

* NOTE: Grey shaded entries are sites from the West Coast region that were excluded from the evaluation of ensemble lead time and high-flow threshold exceedance performance.