Seasonal Hydrological Ensemble Forecasts for Australia using AWRA-L – Hindcast Verification Report

Christopher A. Pickett-Heaps & Elisabeth Vogel

July 2022
Seasonal Hydrological Ensemble Forecasts for Australia using AWRA-L – Hindcast Verification Report

Christopher A. Pickett-Heaps & Elisabeth Vogel

Bureau Research Report No. 065

July 2022
Enquiries should be addressed to:

Christopher A. Pickett-Heaps
Bureau of Meteorology
GPO Box 1289, Melbourne
Victoria 3001, Australia

christopher.pickett-heaps@bom.gov.au:

Copyright and Disclaimer

© 2022 Bureau of Meteorology. To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of the Bureau of Meteorology.

The Bureau of Meteorology advise that the information contained in this publication comprises general statements based on scientific research. The reader is advised and needs to be aware that such information may be incomplete or unable to be used in any specific situation. No reliance or actions must therefore be made on that information without seeking prior expert professional, scientific and technical advice. To the extent permitted by law and the Bureau of Meteorology (including each of its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.
Contents

Executive Summary ........................................................................................................................................ 1

1. Introduction ............................................................................................................................................. 2

2. The Seasonal forecasting system ............................................................................................................. 3
   2.1 Overview ............................................................................................................................................... 3
   2.2 Climate inputs – ACCESS-S2 .............................................................................................................. 4
   2.3 Hydrological modelling system – AWRA-L .......................................................................................... 6
   2.4 AWRA-L seasonal forecast variables ................................................................................................. 7
   2.5 Data assimilation ................................................................................................................................. 8
   2.6 Forecast post-processing ...................................................................................................................... 8
       2.6.1 Temporal and spatial (regional) aggregation .................................................................................. 8
       2.6.2 Absolute and relative forecasts .................................................................................................... 8

3. Forecast evaluation approach .................................................................................................................... 9
   3.1 Hindcast generation .............................................................................................................................. 9
       3.1.1 ACCESS-S2 hindcast data ............................................................................................................. 9
   3.2 AWRA-L hindcasts ............................................................................................................................... 10
   3.3 Reference data for forecast assessment .............................................................................................. 11
   3.4 Forecast skill metrics ............................................................................................................................ 13
       3.4.1 Deterministic skill metrics ......................................................................................................... 13
       3.4.2 Skill metrics for ensemble forecasts ........................................................................................... 14
   3.5 Skill metrics at grid cell and regional scale ......................................................................................... 15
       3.5.1 Evaluation of gridded forecasts .................................................................................................... 15
       3.5.2 Evaluation of forecasts for Geofabric river regions ..................................................................... 16

4. Results and discussion ............................................................................................................................... 17
   4.1 Performance at gridded scale ............................................................................................................... 17
       4.1.1 Deterministic skill ....................................................................................................................... 17
       4.1.2 Probabilistic skill ....................................................................................................................... 19
   4.2 Performance at Geofabric river regions scale .................................................................................... 20
       4.2.1 Continuous ranked probability skill score (CRPSS) ................................................................... 20

5. Discussion ............................................................................................................................................... 22

6. Conclusions ............................................................................................................................................. 23

7. List of references ..................................................................................................................................... 25

8. Supplementary Material ......................................................................................................................... 27
   8.1 Seasonal Skill Maps ............................................................................................................................. 27
       8.1.1 Relative Bias ............................................................................................................................... 27
       8.1.2 Significance in model bias .......................................................................................................... 29
       8.1.3 Anomaly Correlation ............................................................................................................... 30
       8.1.4 CRPS-S ..................................................................................................................................... 32
   8.2 Monthly skill maps ............................................................................................................................... 35
       8.2.1 Soil moisture .............................................................................................................................. 35
       8.2.2 Actual Evapotranspiration ......................................................................................................... 36
       8.2.3 Runoff ....................................................................................................................................... 37
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.3</td>
<td>Reliability (p-value)</td>
<td>39</td>
</tr>
<tr>
<td>8.4</td>
<td>CRPS-S across River Regions</td>
<td>41</td>
</tr>
<tr>
<td>8.4.1</td>
<td>Soil moisture</td>
<td>41</td>
</tr>
<tr>
<td>8.4.2</td>
<td>Actual Evapotranspiration</td>
<td>42</td>
</tr>
</tbody>
</table>
List of Figures

Figure 1: Overview of forecasting system components.................................................. 4
Figure 2: Conceptual AWRA-L v6 grid cell with key water stores and fluxes............... 6
Figure 3: Relative bias of the hindcast compared to the historical reference (in % of mean), at 1-, 2- and 3-months lead time. In case of runoff, an adjustment factor of 1 mm was added to the denominator to avoid unrealistically high values in regions of zero runoff (see Section 3.4.1 for details). ........................................................................... 18
Figure 4: Anomaly correlation at 1-, 2- and 3-months lead time. ............................. 19
Figure 5: Continuous Ranked Probability Skill Score at 1-, 2- and 3-months lead time. .................................................................................................................................. 20
Figure 6: Boxplots of runoff CPRPSS across all 219 Geofabric River Regions for each forecast month. Lead time refers to the lead-time for which the forecast month occurs. ................................................................. 21
EXECUTIVE SUMMARY

The Bureau of Meteorology provides a suite of national gridded hydrological products through the Australian Water Outlook (awo.bom.gov.au). These products underpin services for: (1) current status and historical analysis, (2) seasonal forecasts issued every month and (3) decadal projections out to 2100.

The seasonal forecasting service provides monthly forecasts out to 3 months of key hydrological variables: Root-zone soil moisture, actual evapotranspiration and surface runoff. Forecasts are provided on a ~5km national grid. In addition, a forecast is provided for 219 river basins across Australia, state jurisdictions and nationally across Australia. Forecasts are probabilistic, represented by 99 individual ensembles that relate directly to input climate forcing from a seasonal climate prediction model. The seasonal (3-months ahead) forecasts are issued at the start of the month and available to the Australian public. Key modelling systems supporting seasonal hydrological forecasts are the national surface hydrological model AWRA-L and the ACCESS-S (version S2) operational seasonal climate forecasting system.

This report focuses on an assessment of forecast skill of the AWO seasonal forecasting service. An assessment of forecast skill is based on a series of AWRA-L hindcast simulations using calibrated ACCESS-S2 hindcasts as climate input forcing over a hindcast period of 1981-2017. Calibrated hindcasts refers to the downscaling of ACCESS-S2 hindcasts from a 60km spatial resolution to a 5km resolution via quantile-quantile mapping method. These hindcasts are compared against historical simulations of AWRA-L forced with ‘observed’ climate grids over the same hindcast period. A range of metrics are then calculated to assess forecast performance. This assessment builds on previous work presented in Vogel et al (2021) using ACCESS-S1. Root-zone soil-moisture has positive skill for the 1st forecast month of every month of the year for much of the country. Positive forecast skill is maintained for subsequent forecast months during certain times of the year (e.g. winter in southern Australia). Forecast skill for actual evapotranspiration (ET) is similar to that of soil-moisture. Forecast skill for run-off is reduced compared to soil moisture. However, runoff forecast skill is maintained year-round for the 1st forecast month in non-arid regions and mountainous regions that provide river inflows into the major water storages across the country.
1. INTRODUCTION

Hydrological extremes, such as droughts or floods, can have a devastating impact on many aspects of human societies and the natural environment (IPCC, 2022). A seasonal forecasting service (1- to 3-months in advance) will help adapt to and increase resilience towards hydroclimatic variability and extremes, enabling preparation for potentially harmful events in advance. Seasonal forecasts of key hydrological variables will support improved decision making in many sectors, including water management, agriculture, energy production, emergency services and infrastructure.

The Bureau of Meteorology (“The Bureau”) released the Australian Water Outlook (AWO: https://awo.bom.gov.au) in October 2021. The AWO includes a new national hydrological forecasting and projections service, in addition to a pre-existing historical monitoring service released in 2016 and now repackaged as part of the AWO. The AWO service exists alongside other hydrological forecasting and monitoring services:

- Site-based Flood Forecasting Service
- Catchment based Seven-Day Streamflow forecasting service (updated daily)
- Catchment based Seasonal Streamflow Forecasting Service (updated monthly)
- Catchment based Hydrologic Reference Stations (updated every few years)

The AWO forecasting service consists of a high-resolution national seasonal ensemble forecasting system for soil moisture, evapotranspiration and runoff across Australia. The system applies the AWRA-L gridded surface water balance model forced with calibrated (statistically downscaled) seasonal climate ensemble forecasts from The Bureau’s ACCESS-S (Australian Community Climate and Earth-System Simulator – Seasonal) forecasting system. This report describes the AWO seasonal hydrological forecasting system, its components and presents a comprehensive evaluation of forecast performance across Australia.

Section 2 of this report presents an overview of the seasonal forecasting system and its underlying components. Section 3 outlines the performance evaluation approach, including the hindcast generation, benchmarking data and skill metrics used. In Section 4, the results of the performance assessment are presented and discussed, followed by a discussion and conclusions in Sections 5 and 6 respectively.
2. THE SEASONAL FORECASTING SYSTEM

2.1 Overview

Seasonal hydrological ensemble forecasts are generated by forcing The Bureau’s gridded surface hydrological model (Australian Water Resource Assessment-Landscape model: AWRA-L v6.1, Frost et al. 2018 (a,b)) with calibrated climate input forcing from the ACCESS-S climate forecasting system. Initial states of the hydrological modelling system are updated daily and adjusted through assimilation of satellite-based soil moisture. Following a forecast run at a daily time-step out to 195 days, forecast ensembles are aggregated to a monthly time-step out to 6 months and spatial aggregation is applied across different sets of spatial regions (e.g. the Geofabric V3.2 river regions1, state jurisdictions). Ensemble statistics are finally computed at the AWRA-L grid-scale and spatial regions for display on the AWO web interface.

Figure 1 presents a schematic of the seasonal hydrological forecasting system, including the four main components:

- The climate forecasting system ACCESS-S
- The hydrological modelling system, AWRA-L
- A data assimilation suite
- Spatial and temporal aggregation of forecasts for the web interface

Each of the components is described in the following sub-sections.

---

1 The Geofabric River Regions (V3.2) are a set of hydrological regions defined for regional-scale reporting (hydrological assessments) and modelling, for which detailed metadata is available.
2.2 Climate inputs – ACCESS-S2

ACCESS-S is The Bureau’s sub-seasonal to seasonal climate forecasting system and consists of a coupled atmosphere-ocean-land model, a data assimilation system, and an ensemble generation system. The latest version of ACCESS-S, ACCESS-S2, was deployed to operations on 20th October 2021, replacing ACCESS-S1 (Hudson et al., 2017). ACCESS-S2 includes a locally developed data assimilation system for the ocean and improvements to the land surface initialisation for soil moisture. A detailed description of ACCESS-S and its components can be found via Hudson et al. and http://www.bom.gov.au/research/projects/ACCESS-S/.

ACCESS-S2 is an ensemble seasonal climate forecasting model. Forecast ensembles represent a set of possible forecast outcomes with an assumption of equal likelihood. A forecast for the probability of a given event can then be derived from the ensemble forecasts. A reliable forecast ensures that the eventual observed outcome is statistically indistinguishable from any of the original ensemble members.

Note: A deterministic forecast is a single forecast outcome. If it is generated using a single set of input climate forcings, uncertainty of the deterministic forecast is unknown. However, a deterministic forecast may instead be generated from the mean/median (50% percentile) of a forecast ensemble, and thus the uncertainty of this deterministic forecast may be estimated.

ACCESS-S2 real-time forecasts are generated daily, consisting of 11 ensemble members (the number of which is limited due to computational constraints) and extending out to 217 days. A far greater number of ensemble members are required for acceptable risk evaluation. Consequently, a super ensemble of 99 ensembles is prepared, consisting of ensembles from a 9-
day period starting from the current daily forecast (9 days x 11 ensembles per day). Ensembles from previous daily forecasts are referred to as “lagged ensembles”. For example, a forecast for 1 Jan 2022 would consist of the following ensemble members:

- 1 Jan 2022 – 11 ensembles
- 31 Dec 2021 – 11 ensembles
- 30 Dec 2021 – 11 ensembles
- …
- 24 Dec 2021 – 11 ensembles

Whilst these ensembles are lagged by a different number of days, data processing is applied such that forecasts from all ensemble members start on the same day (e.g. in the example above on 1 Jan 2022). A large number of ensembles (99) increases the likelihood that the probability of extreme events is captured by the probabilistic forecast. However, lagged ensembles imply decreasing forecast skill with increasing lag and suggests a non-uniform weighting-scheme could be applied to the 99-member forecast ensemble. To date, such a weighting scheme has not been developed, and consequently all 99 ensembles are assumed to have equal weight.

Raw ACCESS-S2 atmospheric outputs for each ensemble are available at 60km x 60km spatial resolution and a daily time-step out to 6 months. However, many downstream applications require a higher spatial resolution (e.g. AWRA-L operates on a 5km grid). Raw ACCESS-S2 outputs are calibrated (or downscaled and bias-corrected) towards representing the distribution of higher resolution observational datasets (e.g. the Australian Gridded Climate Dataset, Jones et al., 2009; http://www.bom.gov.au/climate/maps) using quantile-quantile mapping (BoM, 2019). The central idea of quantile-quantile mapping is to match the quantile of an input variable (ACCESS-S2 forecast) at 60km resolution with an equivalent quantile of observations (e.g. AWAP) at 5km resolution. The desired objective of quantile-quantile mapping is two-fold:

- Perform a bias-correction to account for (a) biases in the forecast and (b) topographical effects where the mean differs across the ACCESS-S 60km grid-cell. Here the 50th quantile at 5km resolution is very likely to be different to the equivalent at 60km resolution.

- Account for local extremes at either tail of a forecast distribution that are likely to be quite different at a 5km resolution relative to a 60km resolution

The definition of this mapping is achieved through analysis of historical gridded climatology. Further details are available in BoM (2019).

The calibrated output for each ACCESS-S2 forecast ensemble are available on a 5km x 5km grid resolution and are used as input forcing for the AWRA-L landscape water balance model.
2.3 Hydrological modelling system – AWRA-L

The Australian Water Resource Assessment model (AWRA-L) is a national, grid-based land surface water balance model co-developed by The Bureau and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). AWRA-L is run at a daily time-step and simulates:

- Daily variation in hydrological fluxes (runoff, actual and potential evapotranspiration, deep drainage)
- Daily evolution of hydrological stores (soil moisture and groundwater)

Soil moisture stores are represented as three soil layers (AWRA-L model states) defined on a spatial grid of 5km resolution as follows (Figure 2):

- Upper soil moisture (s0): 0-0.1m
- Lower soil moisture (ss): 0.1-1m
  - Root-zone soil moisture (s0 + ss): 0-1m
- Deep soil moisture (sd): 1-6m

This spatial-temporal specification is consistent with the climate input forcing datasets applied to AWRA-L. Of interest to many users is the root-zone soil moisture.

*Figure 2: Conceptual AWRA-L v6 grid cell with key water stores and fluxes*
AWRA-L model version v6.1\(^2\) is used in the AWO seasonal hydrological forecasting system. AWRA-L is calibrated to observational datasets, including streamflow, remote-sensed soil moisture and evapotranspiration. The performance of the model was comprehensively evaluated against hydrological observations, including gauged streamflow, in-situ measurements of soil moisture, groundwater recharge data, and flux tower-based evapotranspiration (Frost and Wright, 2018). Details on the model structure and calibration can be found in Frost et al. (2018); a comprehensive description of the model evaluation is presented in Frost and Wright (2018).

An AWRA-L forecast run requires a set of initial states. These states are generated by forcing the AWRA-L model with a single set of “historical” analysed climate grids (source: Australian Grided Climate Data - AGCD). Initial states are updated daily and then adjusted via data assimilation of satellite data (section 2.5). The initial states provide a “snapshot” of the current hydrological conditions across Australia.

To generate a seasonal forecast, AWRA-L is forced with calibrated ACCESS-S2 forecasts of rainfall, daily minimum and maximum temperature and solar radiation, and a daily-varying climatology of wind speed on a 5km grid (see Table 1). AWRA-L uses each ACCESS-S2 forecast ensemble member as a separate input forcing dataset, in contrast to historical simulations where only a single set of climate forcings are used. The AWRA-L ensembles together generate a probabilistic forecast of hydrological variables.

**Table 1: Climate forcings used in real-time AWRA-L seasonal forecasts**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily total precipitation (mm)</td>
<td></td>
</tr>
<tr>
<td>Daily minimum temperature (°C)</td>
<td></td>
</tr>
<tr>
<td>Daily maximum temperature (°C)</td>
<td></td>
</tr>
<tr>
<td>Incoming short-wave solar radiation (MJ.m(^{-2}))</td>
<td>Obtained from calibrated ACCESS-S2 forecasts.</td>
</tr>
<tr>
<td>Wind speed (m.s(^{-1}))</td>
<td>A daily-varying wind speed climatology is used which is a default dataset in AWRA-L v6. It is obtained by aggregating wind speed observations used in AWRA-L over the years 1975-2017.</td>
</tr>
</tbody>
</table>

2.4 **AWRA-L seasonal forecast variables**

Seasonal forecasts out to 3 months for the following AWRA-L variables are generated:

- Root-zone soil moisture (the sum of s0 and ss, with a soil depth of 0-1m)
- Surface runoff
- Actual evapotranspiration

\(^2\) AWRA-L v6.1 is a minor update to v6.0 and represents an updated model parameter dataset. All references to AWRA-L v6 are applicable to V6.1.
2.5 Data assimilation

The AWO seasonal forecasting suite applies continental-scale data assimilation developed in collaboration with the Australian National University. The data assimilation system ingests remote-sensed top-layer soil moisture data from the Advanced Scatterometer (ASCAT) and Soil Moisture Active Passive (SMAP) missions and updates the top-layer soil moisture initial states in AWRA-L using triple collocation. The scientific development of the data assimilation system is described in Tian et al. (2021), with an evaluation provided in Bahramian et al. (in prep).

2.6 Forecast post-processing

2.6.1 Temporal and spatial (regional) aggregation

AWRA-L real-time seasonal forecasts are provided to the public at a monthly time-resolution, requiring the temporal aggregation of daily output to monthly estimates. Temporal aggregation is applied as an aggregated sum for flux variables (e.g. Runoff, evapotranspiration) and mean estimate for state variables (i.e., soil moisture).

The spatial representation of the AWRA-L seasonal forecasts is provided as:

- Gridded monthly forecasts at 5km resolution – gridded forecasts are presented as maps to the public.
- Spatially aggregated forecasts for 219 Geofabric river regions (version 3.2, http://www.bom.gov.au/water/geofabric/about.shtml) – spatially aggregated regional forecasts are presented as a box-plot timeseries to the public where the box-plot range represents forecast uncertainty (i.e. forecast ensemble range).

Spatial aggregation is applied to each forecast ensemble member. Spatial aggregation is achieved by calculating a weighted average across all AWRA ~5km grid cells that lie completely or partially within a defined region (e.g. a river basin). Weights reflect the area-fraction of each cell relative to the spatial-region total area (as defined on the 5km grid). Spatial aggregation is applied following temporal aggregation.

2.6.2 Absolute and relative forecasts

AWRA-L seasonal forecasts (equivalent to historical simulations) are provided as two forecast types:

- Absolute forecasts in physical units
- Relative forecasts in percentile units

**Absolute forecasts**: Absolute forecasts are provided in physical units, i.e. mm (for runoff, actual evapotranspiration) or %–full relative to the soil’s available water holding capacity (for root-zone soil moisture). Forecasts provided in physical units are required by many down-stream applications. However, a comparison between forecasts in physical units across space (wet and
dry regions) and time (different seasons) is difficult due to inconsistent (i.e. non-constant) expected values (a different climatology exists depending on the region and/or season).

**Relative forecasts:** Presenting forecasts in percentile units removes the spatial and temporal inconsistency in absolute forecasts, allowing for a comparison between different spatial regions and/or forecast time-steps. Converting to percentile units requires determining the percentile of a forecast ensemble statistic relative to a defined historical reference (or climatology). This allows any user to easily identify regions (or time-steps) where a forecast variable is likely to be near-average, significantly above- or below-average or extreme. The mapping between physical units and percentile units is achieved by calculating the percentile value of a forecast for a given month and variable by comparison against the historical distribution for the same month and variable over the period 1911-2017.

3. **FORECAST EVALUATION APPROACH**

The performance of seasonal AWRA-L forecasts was assessed using retrospective forecasts from ACCESS-S2 (called hindcasts), that are then paired with historical simulations of the same period (individual hindcast dates). This pairing allows for a comparative and forecast skill assessment, generally applied at each month of the year and each forecast lead-time. Hindcasts are generated in the same way as real-time forecasts, although the exact configuration may be different due to computational constraints (e.g. number of ensembles). This evaluation approach is the same as that applied to hindcasts generated using ACCESS-S1 (see Vogel et al., 2021), but has been updated here to reflect the adoption of the ACCESS-S2 system by The Bureau.

The following sections describe the configuration of the AWRA-L hindcasts as well as the performance metrics used to assess the hindcasts.

3.1 **Hindcast generation**

3.1.1 **ACCESS-S2 hindcast data**

*Hindcast period, temporal resolution and post-processing:*

The ACCESS-S2 hindcast period is 1981-2017. This dataset spans 38 years, consisting of a hindcast generated on the 1st of every month and extending forward seven months. While ACCESS-S itself is run at an hourly time-step, the S2 hindcast output consists of aggregated daily totals on a 60km resolution grid.

As with real-time forecasts, the hindcast data are post-processed from a 60km resolution grid to a 5km resolution grid. This post-processing is achieved via a quantile-quantile mapping model (see Section 2.2).

*Hindcast ensembles:*

---

3 The full ACCESS-S2 hindcast is 1980-2018. However, for various reasons a hindcast period of 1981-2017 has been used for the AWRA-L hindcast verification analysis.
Equivalent to the operational ACCESS-S2 forecasts, the hindcasts are based on lagged ensembles. From a given hindcast date (1st of the month), a set of 3 ensembles are generated for 9 consecutive days extending back from the given hindcast date, resulting in an ensemble of 27 ensemble members.

This lagged ensemble configuration closely replicates the configuration used for real-time forecasting, noting the ensemble size of 27 ensemble members is smaller than a 99 ensemble-member real-time forecast (Section 2.2). Computational constraints limit the number of hindcast ensembles; Over a period of 38 years, 444 individual hindcasts were computed in less than 12 months.

### 3.2 AWRA-L hindcasts

#### Hindcast configuration

AWRA-L hindcasts are generated for:

- A period of 1981-2017
- Hindcast start-dates on the first of each month
- Monthly lead-times out to six months
- Ensemble size: 27

**Verification protocol and limitations:**

- Assimilation of satellite soil-moisture (section 2.5) was not included in the hindcast generation and associated historical AWRA-L simulation required for verification (section 3.3). Satellite data were not available for nearly the entire hindcast period. Bahramian et al. (in prep) provides an assessment of the benefit of assimilation of satellite soil-moisture to improve the quality of the AWRA-L initial-states. Consequently, it is likely that data assimilation will further improve forecast skill.

- Deterministic hindcasts were generated from the AWRA-L ensemble forecasts by taking the mean ensemble of each hindcast.

- For experimental purposes, hindcasts of monthly lead-times out to 6 months are available. However, due to limits in forecast skill, forecasts published on the AWO are only for the first 3 months. Consequently, verification analysis provided in this report is limited to 3 months.

The ACCESS-S2 climate input fields for the AWRA-L hindcast match those required for historical simulations. These input fields, described in Table 2, are obtained from ACCESS-S2 (except solar radiation for certain years that were unavailable and wind speed where a climatology was used) and exist at the daily time scale and 5km resolution.
Table 2: Climate forcings used in seasonal AWRA-L hindcasts

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily total precipitation (mm)</td>
<td>Obtained from calibrated ACCESS-S2 hindcasts.</td>
</tr>
<tr>
<td>Daily minimum temperature (°C)</td>
<td>For the years 1990-2016 calibrated solar radiation outputs from ACCESS-S2 hindcasts are used.</td>
</tr>
<tr>
<td>Daily maximum temperature (°C)</td>
<td>From 1981-1989 as well as 2017, outputs from ACCESS-S2 are unavailable. For these time periods, a daily-varying climatology is used as climate forcing.</td>
</tr>
<tr>
<td>Incoming short-wave solar radiation (MJ.m(^{-2}))</td>
<td>The daily-varying solar radiation climatology is a default dataset in AWRA-L v6 and is obtained by aggregating solar radiation observations used in AWRA-L over the years 1990-2017.</td>
</tr>
<tr>
<td>Wind speed (m.s(^{-1}))</td>
<td>A daily-varying wind speed climatology is used which is a default dataset in AWRA-L v6. It is obtained by aggregating wind speed observations used in AWRA-L over the years 1975-2017.</td>
</tr>
</tbody>
</table>

**AWRA-L hindcast output variables**

Seasonal hindcasts were generated to assess the forecast skill of the three AWRA-L seasonal forecast variables (section 2.4) in both physical units (mm) and percentile units (section 2.6.2):

- Root-zone soil moisture
- Surface runoff
- Actual evapotranspiration

The same temporal and spatial aggregation is applied (see Section 2.6.1) and hindcast skill assessment is provided both at grid-cell scale as well as regional scale for the Geofabric River Regions.

### 3.3 Reference data for forecast assessment

A reference dataset is required for a forecast skill assessment and serves two purposes:

- The reference dataset provides data (or “observations”) for direct comparison to forecast estimates. The subsequent forecast-observation pairs may be used to estimate forecast skill, typically by calculating skill metrics.
- The reference dataset provides a benchmark to which a forecast system may be compared. A benchmark is typically a historical/climatological reference for a given time (month) of the year but could also be a previous version of a particular forecasting system.
Application of reference data for verification:

- Historical simulations are treated as “observations” for the purpose of verification but are not true observations of a physical quantity.

- Both hindcast simulations (using ACCESS-S climate forcing) and historical simulations (reference data) are based on the same AWRA-L parameter dataset (v6.1), a single parameter dataset applied to the entire continent. Any impact arising from an overlap between (1) the time-period of the AWRA-L calibration run (1981-2022 + 50-year spin-up) and (2) the AWRA-L hindcast period is expected to be minimal. In addition, differences between hindcast simulations and historical simulations can be attributed to differences in climate forcing (as opposed to approximations arising from AWRA-L model calibration).

- A point-based reference dataset (e.g. observations of gauged streamflow, soil-moisture) could be used as a reference dataset. However, the intention of forecast verification in this report is to quantify the skill of AWRA-L seasonal forecasts using ACCESS-S2 climate forcing relative to simulations using climate gridded forcing (i.e. AGCD). This setup has the advantage that the quantities being compared are the same (e.g. runoff) and the spatial representation (5km grid square) is the same. Conversely, gauged streamflow is not the same physical quantity as ‘runoff’, which is a conceptual variable that cannot be directly measured (but likely to be correlated to observable quantities such as gauged streamflow). Similarly, point-based soil-moisture measurements do not have the same spatial representation as an AWRA-L 5km grid-cell estimate of soil-moisture.

Forecast performance may be assessed in isolation, after creating a dataset of forecast-observation pairs across the entire AWRA-L hindcast period. However, it is also common to assess forecast performance in relation to a benchmark with its own skill level (calculated using the same forecast skill calculation).

A typical benchmark is a historical climatology with a time-invariant expected value and uncertainty. Forecast performance then becomes an assessment of improvement in forecast skill (or lack thereof) compared to a benchmark. Conceptually, an increase in skill usually results from a reduction in the uncertainty of a particular forecast outcome compared to a given benchmark (e.g., a historical climatology) while ensuring the forecast remains reliable (section 3.4.2). A consequence of this is a forecast system exhibits reduced bias (section 3.4.1) by reproducing the inter-annual variability evident in an observed dataset.

The benchmark used in the evaluation of AWRA-L seasonal forecasts is a historical reference (or climatology) for 1981-2017 generated from the same AWRA-L historical simulations mentioned above. These historical simulations were generated using AWRA-L v6.1 and “observed” climate grids as input forcing. Input forcing grids are AGCD grids of precipitation and minimum and maximum daily temperature, satellite-based solar radiation, and interpolated station-based surface wind fields (see Table 3 for data sources). These climate grids are defined on the same 5km grid across Australia as that used for AWRA-L.
Table 3: Climate forcings and data sources for AWRA-L reference simulation

<table>
<thead>
<tr>
<th>Variable</th>
<th>Data source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Daily total precipitation (mm)</td>
<td>Australian Gridded Climate Data (AGCD) (Jones et al., 2009)</td>
</tr>
<tr>
<td>Daily minimum temperature (°C)</td>
<td></td>
</tr>
<tr>
<td>Daily maximum temperature (°C)</td>
<td></td>
</tr>
<tr>
<td>Incoming short-wave solar radiation (MJ.m⁻²)</td>
<td>Dynamic solar radiation based on Grant et al. (2008)</td>
</tr>
<tr>
<td>Wind speed (m.s⁻¹)</td>
<td>Dynamic wind speeds based on McVicar et al. (2008)</td>
</tr>
</tbody>
</table>

3.4 Forecast skill metrics

Forecast skill is assessed using a range of skill metrics which allow for a quantitative assessment (either in isolation or in relation to a benchmark). The following metrics are used for the evaluation of the AWRA-L seasonal hindcast:

- **Relative bias**: mean difference of the forecasts compared to the observations.
- **Anomaly correlation**: a measure of how well the forecast predict the variability around the seasonal mean.
- **Continuous Ranked Probability Skill Score (CRPSS)**: A widely used skill score summarising the accuracy of probabilistic forecasts.
- **Forecast reliability**: A metric capturing how well the uncertainty in forecasts captures the observed variability.

Note: skill metrics can be categorised as ‘deterministic’ or ‘ensemble-based’. Deterministic metrics can only be applied to a single deterministic forecast and do not account for ensemble spread in their mathematical formulation. For ensemble forecasts, deterministic metrics are applied to the mean/median of the ensemble forecasts (e.g. correlation) Ensemble-based metrics account for ensemble spread.

3.4.1 Deterministic skill metrics

**Relative bias**

Relative bias is the expression of model bias as a proportion of the mean observed value.

The relative bias can become unrealistically large for areas where the mean value is close to zero (i.e. leading to division by close-to-zero values). This is particularly the case for regions and months of zero runoff, leading to inflated relative bias values for runoff in dry inland regions. This was addressed by adding a small value (1 mm) to the denominator (Vogel et al. 2021). Hence, the relative bias was calculated as:
\[
Bias_{rel(l)} = \frac{\sum_{i=1}^{N} y_i(l) - x_i}{\epsilon + \sum_{i=1}^{N} x_i}
\]

where \( i = 1, \ldots, N \) are the monthly time steps, \( l \) is the lead time, \( x_i \) is the historical simulation using observed climate inputs, and \( y_i(l) \) is the forecast for the time step \( i \) at lead time \( l \) and \( \epsilon \) is the adjustment factor to avoid inflated relative biases when the mean is close to zero. The adjustment factor was used only for computing bias in runoff and set to 1mm in this case; it was set to 0 for other variables.

**Anomaly correlation**

The anomaly correlation was calculated from the forecast ensemble mean in the following steps:

1. Calculate the climatology over 1981-2017 (i.e. mean values for each month) for the hindcasts and historical reference (i.e. the climatology for the hindcasts and historical reference are computed separately).
2. Subtract the monthly climatology (over 1981-2017) from the forecast ensemble mean and the historical reference, respectively.
3. Subsequently, calculate the Pearson correlation coefficient between forecast anomalies and historical reference anomalies.

A correlation based on anomalies is applied to remove the effect of the climatological seasonal cycle on the correlation coefficient. Anomaly correlations focus on the ability to reproduce deviations from the mean seasonal cycle (i.e. accurately predicting inter-annual variability: wetter- or drier-than-average conditions at a particular time of year).

**3.4.2 Skill metrics for ensemble forecasts**

Ensemble forecast skill metrics consider the ensemble range (or uncertainty) of a forecast.

**Continuous Ranked Probability Skill Score (CRPSS)**

The Continuous Ranked Probability Score (CRPS) (Hersbach et al. 2000) relates to the accuracy of forecast, where a forecast cumulative distribution is compared to an (observed) empirical distribution. CRPS forecast skill is usually expressed as a skill-score (Continuous Ranked Probability Skill Score, CRPSS), calculated as follows

\[
CRPSS = 1 - \frac{CRPS_f}{CRPS_r}
\]

CRPS\(_f\): CRPS of a forecast
CRPS\(_r\): CRPS of a reference (benchmark)
Consequently, a skill-score is unitless:

- Skill-scores > 0 indicate an improvement in forecast skill compared to a benchmark
  - Implies increased forecast accuracy (reduced bias) and a likely reduction in forecast uncertainty compared to that of a benchmark

- Skill-scores ~ 0 indicate equivalent forecast skill compared to a benchmark
  - Implies no increase in forecast accuracy and very likely no reduction in forecast uncertainty

- Skill-scores < 0 indicate a decrease in forecast skill compared to a benchmark
  - Implies reduced forecast accuracy (likely increased forecast bias) compared to a benchmark. Any reduction forecast uncertainty range compared to a benchmark is inconsequential as the forecast remains inaccurate.

Here we used a climatology forecast as a benchmark for calculating the CRPSS.

**Forecast reliability**

Forecast reliability (Wilks 2011) specifically relates forecast probabilities to observed frequencies for a set of forecast values. Ideally, this relationship is close to 1:1, indicating a close agreement. More generally, forecast reliability relates to forecast ensemble consistency, where ideally an observed outcome is statistically equivalent to (and therefore indistinguishable from) any forecast ensemble.

Forecast reliability is assessed by evaluating rank histograms which visualise the distribution of the ranks of each observation relative to a forecast ensemble over a large sample of forecasts. The ranks of the observations will be uniformly distributed (Wilks, 2011) if a forecasting system is reliable. After calculating the rank of each observation within a forecast ensemble spread, the Kolmogorov-Smirnov (KS) statistical test verifies if the ranks are uniformly distributed. A p-value < 0.05 indicates a high confidence that the distribution of ranks is significantly different from a uniform distribution, indicating the forecast system is not reliable.

### 3.5 Skill metrics at grid cell and regional scale

Forecast skill metrics are calculated for every month a forecast is issued and every monthly lead time for which a forecast is provided. Forecast skills are calculated for both absolute forecasts (physical units) and relative forecasts (percentile units). See section 2.6.2.

#### 3.5.1 Evaluation of gridded forecasts

Forecast skills are calculated for every grid cell across the entire country (on a 5km x 5km grid). Each grid cell of the AWRA-L grid can be considered as an individual data time-series. At a high level, the following steps are followed:
• Monthly aggregation of daily data:
  − Daily hindcast data grids are aggregated to monthly grids
  − Daily “observed” grids are aggregated to monthly grids

• Conversion from physical units (mm) to percentile units
  − Conversion applied to both hindcast data grids and observed grids

• Hindcast data paired with observed grids
  − Hindcast data in physical units paired with observations in physical units
  − Hindcast data in percentile units paired with observations in percentile units

• Forecast metrics calculated with paired datasets – stratified by forecast month and lead time

As mentioned in section 3.3, AWRA-L historical simulations are used in the evaluation of gridded forecasts. These historical simulations are in turn generated using input forcing from the AGCD climate grids. These climate grids are generated through the interpolation of point-based observations, the network of which is very sparse in certain remote areas of Central Australia. Consequently, the quality of interpolated gridded climate data is reduced. These regions are excluded from subsequent skill analysis.

### 3.5.2 Evaluation of forecasts for Geofabric river regions

Forecast skills are also calculated for all Geofabric (geospatial) river regions (219 regions). At a high level, the following steps are followed:

• Monthly aggregation of daily data grids
  − Daily hindcast data grids are aggregated to monthly grids
  − Daily “observed” grids are aggregated to monthly grids

• Time-series of monthly data are generated for each Geofabric region via spatial weighted averaging of gridded data lying within a region
  − Monthly time-series generated for both hindcast data and observed data

• Conversion from physical units (mm) to percentile units for each region time-series
  − Conversion applied to both hindcast data grids and observed grids

• Hindcast data paired with observed grids
  − Hindcast data in physical units paired with observations in physical units
  − Hindcast data in percentile units paired with observations in percentile units

• Forecast metrics calculated with paired datasets – for every forecast month and every lead time separately
4. RESULTS AND DISCUSSION

4.1 Performance at gridded scale

This section includes a discussion of relative bias, anomaly correlation, CRPSS and forecast reliability. The discussion focuses on aggregated annual forecast skill and seasonal forecast skill (Section 8.1) to provide a succinct summary of the performance of the AWRA-L seasonal forecasting system overall. An analysis of forecast skill for individual months is not included in this section as it is largely reflected in the seasonal analysis. However, monthly verification maps (relative bias and CRPSS) that correspond to forecasts issued every month to the public are included in section 8.2. Section 8.3 contains summary maps of forecast reliability.

Many of the figures in this section contain maps of forecast performance across Australia. Regions in western Central Australia are masked out in these maps due to a sparse observation network and consequently reduced quality in the climate reference dataset (see section 3.5.1).

4.1.1 Deterministic skill

**Relative bias**

The relative bias of the forecast outputs for soil-moisture and actual evapotranspiration is close to zero across Australia for all lead times (Figure 3). Curiously, the relative bias for actual evapotranspiration decreases with increasing lead-time whereas one would expect the opposite. However, both the bias magnitude and trend with lead-time are small and were found to be insignificant. Other metrics (e.g. CRPS-S) indicate an expected fall in forecast skill with lead-time (section 4.1.2). While the relative bias values for runoff (in %) are larger compared to those of soil moisture and evapotranspiration, the variability of runoff is also larger, implying forecast bias is likely not statistically significant.

Overall, forecast bias is not significant for all forecast variables. This insignificance demonstrates the acceptable performance of quantile-quantile bias correction applied to climate input forcing (Section 2.2). Significance in bias (both relative and absolute) was identified via bootstrap sampling and an analysis of the resulting distribution in bias. Significance in forecast bias was identified if the resulting confidence interval (2.5 percentile, 97.5 percentile) excluded zero bias. Over much of the country, forecast bias was found to be generally insignificant. Certain areas do show significant bias, but in many cases the actual bias (and confidence interval in bias) is small. Section 8.1.2 includes maps of the confidence interval in the annual mean forecast bias. This finding is similar to Vogel et al 2021 who presents a similar evaluation based on hindcasts using ACCESS-S1 climate forcings (the predecessor of the current climate forecasting suite).

At the seasonal scale (Section 8.1), bias maps show relatively minor variations across the different seasons for soil moisture and evapotranspiration. The bias is also somewhat stable across lead-time. Runoff exhibits somewhat small variations, with positive biases (i.e. wetter conditions) in DJF, MAM and JJA and to a lesser extent in SON).
**Anomaly correlation**

Figure 4 presents the anomaly correlations between the forecast ensemble mean and the historical reference, for each variable and lead time. The results indicate positive anomaly correlations consistent across the continent for soil moisture and evapotranspiration, with particularly high correlations (>0.3) at lead-1 and lead-2, and to some extent at lead-3. For runoff, high anomaly correlation values can be found across important runoff-generating regions at lead-1.

Seasonal-scale anomaly correlation plots are included in Section 8.1.3 and show that for soil moisture and evapotranspiration, the anomaly correlations are high throughout all seasons, with particularly high correlations at longer lead times in JJA and SON. For runoff, anomaly correlations are high at lead-1 for MAM, JJA and SON. At longer lead times, anomaly correlations for runoff are highest for SON, indicating predictability at longer lead times across large parts of Australia in spring.

*Figure 3: Relative bias of the hindcast compared to the historical reference (in % of mean), at 1-, 2- and 3-months lead time. In case of runoff, an adjustment factor of 1 mm was added to the denominator to avoid unrealistically high values in regions of zero runoff (see Section 3.4.1 for details).*
4.1.2 Probabilistic skill

**CRPS Skill-score (CRPSS)**

Figure 5 presents the overall CRPSS across all months of the year, for soil moisture, actual ET and runoff, across lead times 1-3. CRPSS values greater than zero (blue colour) indicate that the forecast performs better than a climatology forecast, whereas values close to zero (white) or negative (red) suggest that the forecast performs similar to or worse than climatology. For all three variables, the forecast skill is positive across most of Australia at lead-1. The positive skill extends into lead-2 and to some extent lead-3 for actual ET and soil moisture depending on the region (eastern Australia), whereas runoff skill decreases rapidly after the first month of the forecast.

These results indicate a potential for forecasting soil moisture and actual ET two months (and in certain regions up to three months ahead). While forecast skill for runoff is reduced (positive skill for lead-1 only), these results are an aggregation across the year. For individual seasons and catchments, we see positive skill for runoff at longer lead times (seasonal/monthly CRPSS results are included in Sections 8.1.4 and 8.2 respectively). For JJA and SON, runoff shows positive skills at longer lead times (up to lead-2 and -3) in important runoff-generating regions such as the Great Dividing Range, large areas in Western Australia and Northern parts of the country. For DJF and MAM, the CRPS skill drops off quicker with lead-time. The reasons this seasonality in forecast skill with lead-time are discussed further in the discussion (section 5).
Forecast reliability

Forecast reliability was assessed for all months of the year and lead-times given the forecasts provided to the public are issued monthly. Summary maps for a selection of months of the year are provided in section 8.3 and are reflective of the reliability of across all months of the year. The AWRA-L seasonal forecasting system generally provides reliable seasonal forecasts for most times of the year (and lead-times) in most regions across Australia. Seasons and/or regions for which forecasts are not reliable are those where seasonal or annual rainfall is low (and consequently little variation in surface hydrology is observed). Such regions include central Australia and the dry season across northern Australia.

4.2 Performance at Geofabric river regions scale

4.2.1 Continuous ranked probability skill score (CRPSS)

This section presents the CRPSS for runoff across all 219 Geofabric River Regions (Figure 6). The results of the regional-scale CRPSS for soil moisture and actual evapotranspiration can be found in Section 8.4.

The CRPSS for runoff shows positive skill for more than 75% of river regions at lead-1 (first month of the forecast) across all forecast months (Figure 6). For some months of the year, the forecast skill is particularly high, with positive skill extending into longer lead times. In July to October, more than 75% of River Regions show positive skill up to lead-2. And for June to October the median CRPSS is positive for all three lead times, indicating that at least 50% of stations have positive skill out to the third month of the forecast.
Figure 6: Boxplots of runoff CPRPSS across all 219 Geofabric River Regions for each forecast month. Lead time refers to the lead-time for which the forecast month occurs.
These results suggest that, although the runoff skill at grid cell scale, averaged over all months of the year (Figure 5) indicates rapidly declining skill after lead-1, we find that for specific regions and seasons, the forecasts can provide skillful runoff forecasts up to two and three months ahead.

NOTE: As with the skill-score maps, the skill-scores plotted in Figure 6 and section 8.4 are for a forecast month when that forecast month occurs at a specified forecast lead-time. For example, September when September is at lead-time L2 implies a forecast issue date of July.

5. DISCUSSION

Positive forecast skill (e.g. CRPS-S > 0.1 together with a forecast that is reliable) indicates forecast products can be expected to provide a benefit to decision making if a decision would otherwise be based on a benchmark product such as a historical reference. Insignificant forecast skill (e.g. CRPS-S: [-0.1, 0.1]) indicates limited benefit compared to using a simple benchmark such as a historical climatology. Negative forecast skill (CRPS-S < -0.1) indicates forecasts should be used with greater caution or eliminated from the decision-making process depending on the nature of the decisions being made. In all cases, any forecast shown as unreliable (section 8.3) should be used with great caution.

As described in the results section (section 4), forecast skill varies in relation to:

- Lead-time
- Time of year (month)
- Physical quantity (e.g., soil-moisture, runoff or actual ET)

In almost any forecasting application, forecast skill always reduces with increasing lead-time due to inherent model inaccuracies and approximations. Climate and hydrological forecasting are no different.

Reduced forecast skill is also apparent at different times of the year, typically during the season preceding the rainy season of a particular region (winter in Southern Australia and summer in Northern Australia). In climate forecasting, this is often referred to as the Autumn Predictability Barrier for Southern Australia and feeds into the lack of forecast skill in hydrological forecasts. A compounding factor is the influence of hydrological initial states (or initial conditions) on predictive skill, which again is reduced during periods preceding the “rainy season” and the period in the year in which the “rising limb” is observed in either soil-moisture or runoff. In contrast, initial states have far greater predictive skill during periods in which the “falling limb” is observed, due in part to hydrological persistence in time.

Another factor influencing reduced forecast skill is whether variability in the expected outcome of a physical quantity (e.g., runoff) is already small. Limited variability typically occurs in the months of reduced rainfall (e.g., summer in southern Australia, dry season in northern Australia) and in arid regions of central Australia. Limited variability can also occur in regions/locations where the observed inter-annual variability around a climatological expected value is small. Any subsequent improvement from a forecast is limited and likely insignificant, resulting in an interpretation of “no improvement in forecast skill”. Regions with large inter-annual variation
offer greater potential for a skilful seasonal forecasting system given greater scope to reduce the uncertainty in an expected outcome. Much of Australia exhibits significant inter-annual variation in water resources due to significant inter-annual climate variability (e.g., arising from the El-Nino Southern Oscillation, Indian Ocean Dipole and Southern Annular Mode). The variation in forecast skill between different physical quantities, particularly root-zone soil moisture and runoff, is primarily due to different factors influencing the estimation of these physical quantities. The skill in forecasting root-zone soil moisture is primarily influenced by initial states of the hydrological model, as opposed to input climate forcing. Forecast skill for root-zone soil moisture (and to some extent evapotranspiration) is clearly superior to runoff, although seasonal variations in skill are still apparent, particularly at longer lead-times. Runoff is instead primarily influenced by forecast precipitation from a climate model. Limited forecasting skill of seasonal rainfall therefore directly impacts forecasting skill of surface runoff. Forecast skill for actual ET appears to be reduced compared to root-zone soil-moisture but improved compared to runoff. This may reflect competing influences from climate input forcings on actual ET and consequently the forecast skill of actual ET relative to other forecast variables. Take note that climatological wind fields instead of dynamic wind fields are included as part of the input climate forcing. Dynamic (time-varying) wind fields may well improve forecast skill for actual ET.

Due to the lack of available satellite soil-moisture data for most of the ACCESS-S2 hindcast period (1981-2017), an analysis of the impact on forecast skill from data assimilation is currently not possible. Satellite data of soil-moisture only became available from ~2016 (CHECK). However, it is worth considering the positive impact data assimilation might have on forecast skill. The benefit of data assimilation is to improve the initial states (soil-moisture stores) of the AWRA-L prior to a forecast run. Ostensibly this suggests DA would improve forecast skill of root-zone soil-moisture. However, data assimilation specifically adjusts the top-layer soil-moisture (s0) which is heavily influenced by input rainfall (in the form of daily rainfall climate grids). An improved initial state of the top-layer soil-moisture layer may in fact improve forecasting skill of runoff.

6. CONCLUSIONS

The Bureau of Meteorology launched a new seasonal hydrological forecasting service in October 2021. The service is one of three services available from The Australian Water Outlook. The forecasting service provides monthly forecasts out to 3 months of key hydrological variables: Root-zone soil moisture, actual evapotranspiration and surface runoff. The forecasts are provided on a 5km national grid as well as for 219 river basins across Australia. Seasonal (3-month) forecasts are issued at the start of the month and available to the Australian public.

The assessment of forecast skill in this report indicates positive skill for root-zone soil moisture for the 1st forecast month (lead-time 1) of every month of the year. Positive forecast skill is maintained for subsequent forecast months (lead-times 2, 3) during and following periods of relatively high seasonal rainfall (the “rainy” seasons either in southern or northern Australia). Overall, forecast skill in actual ET is similar to that of root-zone soil-moisture.
Forecast skill for run-off is somewhat reduced compared to root-zone soil moisture. This is primarily due to limitations in seasonal rainfall forecast skill. However, positive forecast skill is maintained year-round for the 1st forecast month (lead-time 1) in non-arid regions where reasonable rainfall can be expected. In particular, positive skill is maintained in mountainous regions that provide river inflows into the major water storages across the country, for example in South-Eastern Australia and along the Great Dividing Range in Eastern Australia.

Consideration of forecast skill, and particularly the spatial-temporal variations therein, is a key factor to ensure the AWO forecasting service is used appropriately by end-users to make informed decisions. Transparency in forecast quality/reliability allows users to decide when, and when not to use a forecast product. The benefit of using forecasts as an aid to decision making, indicated by forecast skill, is particularly relevant if decision makers were to use a simplified product such as a historical reference in the absence of a forecast. If an end-user were to make a decision without reference to any such simplified product, the use of an available forecast product, despite limited skill, may still provide benefit.

Forecast skill extending out to multiple lead-times will have a significant benefit either directly or indirectly on downstream applications. Overall, forecast skill demonstrates the value of the AWO seasonal forecasting service to decision makers with an interest in forecasts of hydrological quantities such as root-zone soil moisture, actual ET and runoff.
7. LIST OF REFERENCES


8. SUPPLEMENTARY MATERIAL

8.1 Seasonal Skill Maps

8.1.1 Relative Bias

![Seasonal Skill Maps Diagram]

- **DJF**
  - Lead-1: Soil moisture mean: 0.76%
  - Actual evapotranspiration mean: -3.46%
  - Runoff mean: 10.77%
  - Lead-2: mean: 1.30%
  - mean: -2.84%
  - mean: 12.63%
  - Lead-3: mean: 2.34%
  - mean: -1.69%
  - mean: 13.22%

- **MAM**
  - Lead-1: Soil moisture mean: 0.86%
  - Actual evapotranspiration mean: -1.92%
  - Runoff mean: 6.05%
  - Lead-2: mean: 1.74%
  - mean: -1.91%
  - mean: 8.72%
  - Lead-3: mean: 1.55%
  - mean: -2.96%
  - mean: 9.28%
8.1.2 Significance in model bias

Absolute bias (annual mean)

Relative bias (annual mean)

2.5%ile

97.5%ile
8.1.3 Anomaly Correlation

DJF

Lead-1
Soil moisture mean: 0.72
Actual evapotranspiration mean: 0.05
Runoff mean: 0.28

Lead-2
mean: 0.22
mean: 0.35
mean: 0.08

Lead-3
mean: 0.13
mean: 0.23
mean: 0.07

MAM

Lead-1
Soil moisture mean: 0.82
Actual evapotranspiration mean: 0.92
Runoff mean: 0.42

Lead-2
mean: 0.33
mean: 0.50
mean: 0.15

Lead-3
mean: 0.16
mean: 0.17
mean: 0.10
SEASONAL HYDROLOGICAL ENSEMBLE FORECASTS FOR AUSTRALIA USING AWRA-L – HINDCAST VERIFICATION REPORT

JJA

SON

Lead-1

Soil moisture
mean: 0.84

Actual evapotranspiration
mean: 0.91

Runoff
mean: 0.38

Lead-2

mean: 0.42

mean: 0.62

mean: 0.24

Lead-3

mean: 0.25

mean: 0.39

mean: 0.21

Lead-1

Soil moisture
mean: 0.77

Actual evapotranspiration
mean: 0.89

Runoff
mean: 0.46

Lead-2

mean: 0.36

mean: 0.55

mean: 0.32

Lead-3

mean: 0.32

mean: 0.48

mean: 0.31

Anomaly correlation (Pearson-R)

-1.00

-0.60

-0.30

-0.20

-0.10

0.00

0.10

0.20

0.30

0.40

0.50

0.60

0.70

0.80

0.90

1.00
8.1.4 CRPS-S

DJF

- Lead-1: Soil moisture mean: 0.35
- Lead-2: mean: 0.06
- Lead-3: mean: -0.01

MAM

- Lead-1: Soil moisture mean: 0.49
- Lead-2: mean: 0.14
- Lead-3: mean: -0.03

Continuous rank probability skill score

- Actual evapotranspiration
- Runoff
- mean: 0.08
- mean: -0.04
- mean: -0.04
- mean: 0.17

8.2 Monthly skill maps

8.2.1 Soil moisture

Relative bias

CRPSS
8.2.2 Actual Evapotranspiration

Relative bias

CRPSS

[Images of maps showing relative bias and CRPSS values for different months and lead times]
8.2.3 Runoff

Relative bias

CRPSS
8.3 Reliability (p-value)

Jan

Lead-1
Soil moisture mean: 0.54
Actual evapotranspiration mean: 0.32
Runoff mean: 0.56

Lead-2
mean: 0.57
mean: 0.48
mean: 0.56

Lead-3
mean: 0.59
mean: 0.50
mean: 0.59

Apr

Lead-1
Soil moisture mean: 0.45
Actual evapotranspiration mean: 0.37
Runoff mean: 0.48

Lead-2
mean: 0.54
mean: 0.55
mean: 0.59

Lead-3
mean: 0.60
mean: 0.55
mean: 0.61
8.4 CRPS-S across River Regions

8.4.1 Soil moisture
8.4.2 Actual Evapotranspiration

S2:hindcast: Variable etot - phys - crps_ss_rel - Q50
Forecast month (issue_mth = fcst_mth - leadtime)