$$NSE = 1 - \left(\sum_{t=1}^{n} \left(x_{med}(t) - Q_{obs}(t) \right)^2 \Big/ \sum_{t=1}^{n} \left(Q_{obs}(t) - \overline{Q_{obs}(t)} \right)^2 \right), \quad (5)$$

$$Pbias = \left[\sum_{t=1}^{n} \left(x_{med}(t) - Q_{obs}(t) \right) \Big/ \sum_{t=1}^{n} Q_{obs}(t) \right] \cdot 100\% \quad (7)$$

$$\overline{MAD} = \frac{1}{n} \sum_{t=1}^{n} median_i \left| x_i(t) - x_{med}(t) \right| \quad (2)$$

A simple measure of ensemble accuracy is the Containing Ratio (CR) (Xiong and O'Connor, 2008):

$$CR = \frac{1}{n} \sum_{t=1}^{n} I[Q_{obs}(t)] \quad (8)$$

where $I[\bullet]$ is an indicator function as follows:

$$I[Q_{obs}(t)] = \begin{cases} 1, x_{(1)}(t) < Q_{obs}(t) < x_{(z)}(t) \\ 0, otherwise \end{cases} . \quad (9)$$

$I[Q_{obs}(t)]$ equals 1 when the observation falls between the lowest and highest valued ensemble members and $I[Q_{obs}(t)]$ equals 0 when the observation falls outsize the ensemble bounds.

2.4.4 Conditional Statistics

In the previous sections, we presented metrics that compare the simulated discharge values (i.e. median, minimum and maximum of the ensemble) to observed discharge values. In the following section, methods that evaluate probability values from the ensemble for specific discharge events are presented.

We first define $m_i(t)$ as the probability of a simulated streamflow event at a given timestep from the model ensemble, which can take on any of I values $m_1(t), m_2(t) \ldots m_I(t)$ (Wilks, 2006). The corresponding observation ($y_j(t)$) can take on any of J values $y_1(t), y_2(t) \ldots y_J(t)$. In this study, three possible observations (i.e. $J = 3$) are defined: low flow or a discharge value that is less than the 30th percentile of climatology; middle flow or a discharge value that is between the 30th and 70th percentiles of climatology; and high flow or a discharge value that is greater than the 70th percentile of climatology. Climatology is based on the available discharge data at each site (Table 1).

The probability of a simulated streamflow event is derived by computing the percentage of the ensemble members that fall within each flow category at a given timestep. The probability is rounded up to the nearest tenth probability, therefore the probability will fall within one of ten possible probability bins (0-10%, >10%-20%, etc.). At a given timestep, the observation will have a value of 1 ($y_j(t)=1$) for the flow category in which it was observed, and a value of 0 ($y_j(t)=0$) for the flow categories in which it did not occur.

Murphy and Winkler (1987) set up a general framework for forecast verification based on factorization of the joint distribution of forecasts and observations into the calibration-refinement factorization:

$$p(m_i, y_j) = p(y_j|m_i)\,p(m_i);\ \text{i} = 1,\ldots,\text{I};\ \text{j}=1,\ldots,\text{J}. \qquad (10)$$

and the likelihood-base rate factorization:

$$p(m_i, y_j) = p(m_i|y_j)\,p(y_j);\ \text{i} = 1,\ldots,\text{I};\ \text{j}=1,\ldots,\text{J}. \qquad (11)$$

The conditional distribution $p(y_j|m_i)$ in Equation 10 is the more familiar measure of the two and can be plotted on a reliability diagram as a function of the ensemble probability. The ensemble probability is well calibrated if, for a given flow category, the relative frequency of the conditional event equals the ensemble probability (e.g. $p(y=low\ flow|m=0.1) =0.1)$ and when plotted on the reliability diagram, the conditional event will plot along a 1:1 line (Murphy and Winkler, 1987; Murphy and Winkler, 1992; Wilks, 2006). To avoid confusion with the model parameter calibration discussion, hereafter we refer to the calibration of the ensemble probability as reliability.

The relative frequencies of the ensemble probabilities ($p(m_i)$) are plotted as an inset on the reliability diagram to indicate the sharpness, or resolution, of the ensembles (Wilks, 2006). Sharp ensembles will have narrowly distributed probability values where probability occurs most frequently in the extreme probability categories (i.e., 0-10% and >90-100%).

The likelihood distribution ($p(m_i|y_j)$) is a less intuitive measure, but very useful for evaluating how much probability the ensemble gives to the correct flow category compared to other possible categories. For all instances of an observation occurring in a given flow category, the conditional probability for all possible flows is computed: for example, the ensemble probability of a low flow given a low flow observation ($p(m=low\ flow|y=low\ flow)$), the ensemble probability of a middle flow given a low flow observation ($p(m=middle\ flow|y=low$

flow)), and the ensemble probability of a high flow given a low flow observation (*p(m=high flow|y=low flow)*). These likelihood distributions can then be plotted on the discrimination diagram as a function of the ensemble probability. Ensembles are highly accurate if the majority of the ensemble members frequently fall within the flow category observed (in the previous example, this would be the low flow category), resulting in high probabilities for the observed flow category and low probabilities for the remaining flow categories. For such ensembles, the likelihood distributions for the different possible flows will not overlap to a great degree when plotted on the discrimination diagram and they are considered to have good discrimination for that flow category (Murphy and Winkler, 1987; Murphy et al., 1989; Wilks, 2006).