Technical Note: T15-3

RAINFALL STATISTICS – AN EXERCISE:
Choose an appropriate statistic

Dr Robert Patterson
FIEAust, CPSS(3), CAg

Armidale NSW
17th December 2020

RAINFALL STATISTICS – AN EXERCISE: CHOOSE AN APPROPRIATE STATISTIC

1. Introduction

You may ask “What is a water balance model?” That question, answered in the *Water Balance Technical Sheet* is simply a number of simple arithmetical calculations that account for the internal wastewater generation and rainfall inputs to the model and the outputs or returns of the water to the hydrologic cycle as drainage, evapotranspiration, runoff and changes in soil moisture. Some data will be obtained from the wastewater report, other information is sourced from the internet.

Water balance modelling can be performed on various data sets depending upon the data available and the degree of precision provided by the modelling calculations. The water balance model is simply a number of calculations of simple formulae that can be performed by a computer much faster than one could do the same calculations by hand, often seconds compared with hours. However, you have probably head the expression "garbage in - garbage out" meaning that the output of the model (your assessment of the land application area required) can only ever be as good as the data selected for the model. Here lies the catch - do you have a daily time step model using all the historical rainfall recording for your location, and the daily evaporation date, or do you use monthly historical rainfall and evaporation data, or some combination of the two. Daily rainfall may appear more precise but the calculations are based upon previous rainfall records and may not predict future rainfall periods with anything more than a ‘best guess’.

Since rainfall may be seasonal based upon world climatic zone, its occurrence is random, meaning that it does not necessarily rain on every Monday, or that because it rained yesterday it may rain again today. So, for any one year, the rainfall pattern may be entirely different for the day before or the day after of the same date of another year, and that the total annual rainfall last year may be entirely different from a decade ago, or even next year. These differences are well reflected in historical rainfall records. This exercise will help you explore some of these differences.

This difference can be shown in Figure 1 where the three rainfall periods for January 1996, 2006 and 2016 have been graphed on the same axis for the days of the month. There is little similarity of rainfall at any period within those months and even the monthly totals are significantly different. How do you plan for this range of variation?

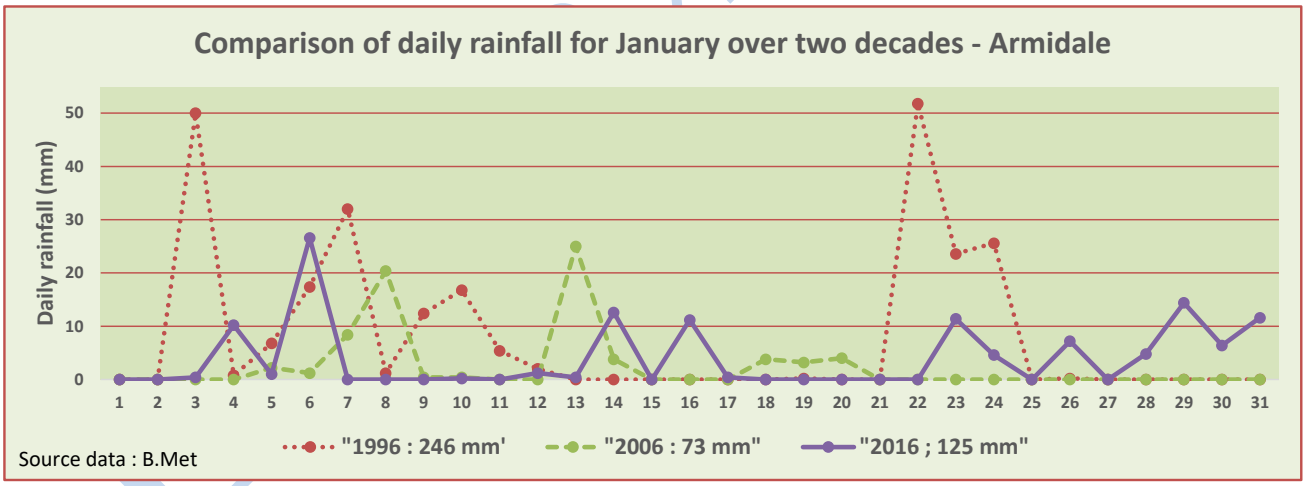

Figure 1 Comparison of three January rainfall periods

2. Rainfall Statistics – an exercise

The Australian Bureau of Meteorology produces statistical summaries for official rainfall recording stations that may or may not include recordings of pan evaporation. The Bureau’s Climate Data On-line (www.bom.gov.au/climate/data) is the first port of call to see just what site is available, closest to our locality of interest. The data derived to construct Figure 1 were sourced from the Bureau’s website.

Open the website and following the instruction for the example of Armidale, where the data from two recording stations need to be merged to provide the full historical model. If one is using a smaller range of years, such as a 30 or 40 year period, there may need to be some correlation of that period to the overall record as will be explained in Section XXX.

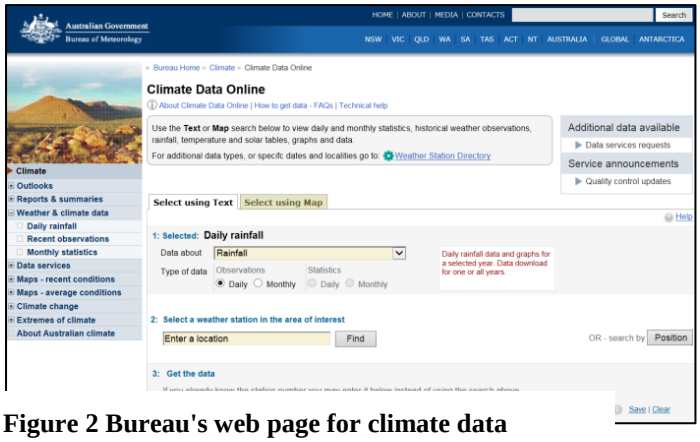

Figure 2 Bureau's web page for climate data

Item 1: you can select rainfall, weather & climate, temperature or solar exposure.

For this exercise choose 'weather & climate' and choose 'monthly' under 'statistics'.

Item 2. Enter the location of interest – in this case choose ARMIDALE, then FIND

Choose Armidale NSW and note the station and surrounding stations that are available. Choose 056037 which then appears under Item 3.

Click 'Get Data'. A new page opens in a new window.

Evaporation data

In the bar under the map is listed 'nearest alternative sites', click the 'All available' and a larger set of tables is opened. Look down the 'Statistics' column for evaporation, and note that evaporation data are available for Armidale 056037. Note that this data set only commenced in 1997, so the period is very short and unlikely to show long term trends or cycles, but is suitable for most wastewater modelling.

Unfortunately the data are not readily available for open pan evaporation for the years of recording so we need to be satisfied with the average pan evaporation data that the Bureau of Meteorology provides on its website.

Rainfall data

Return to the main page and under 'Nearest Bureau Stations', uncheck the box 'Only show open stations' and you will note that there is now access to 056002 Armidale (Radio Station 2AD) at the top of the list. Check this location, get data and note that the period is from 1857 to June 1997 – a much longer period to see longer annual and seasonal patterns in rainfall. However, note there is no 'evaporation' for this station.

Now return to the main "Climate Data Online' webpage and under item 1 select 'Rainfall' and 'Monthly observations. Item 2 remains Armidale, select Armidale under 'Matching towns" leave blank the box "Nearest Bureau Stations' and first select 056002 Armidale (Radio Station 2AD) and a bar will show available data. Select "Get Data".

You can save the spreadsheet simply by selecting "All years of data' in top right hand corner of the open 'Monthly rainfall' sheet. Open the saved spreadsheet in a new Excel™ spreadsheet Return to the "Climate Data Online' webpage' and repeat the process for Stn 56037 (Armidale Tree Group Nursery). Combine both sets of data into one spreadsheet and arrange the data from the earliest to the latest date using the 'sort' facility in the spreadsheet.

A graph of the saved data for Armidale can be plotted, as in Figure 3 that reflects the random nature of the rainfall. A trend line shows a slight decline in annual rainfall over the period of the data set.

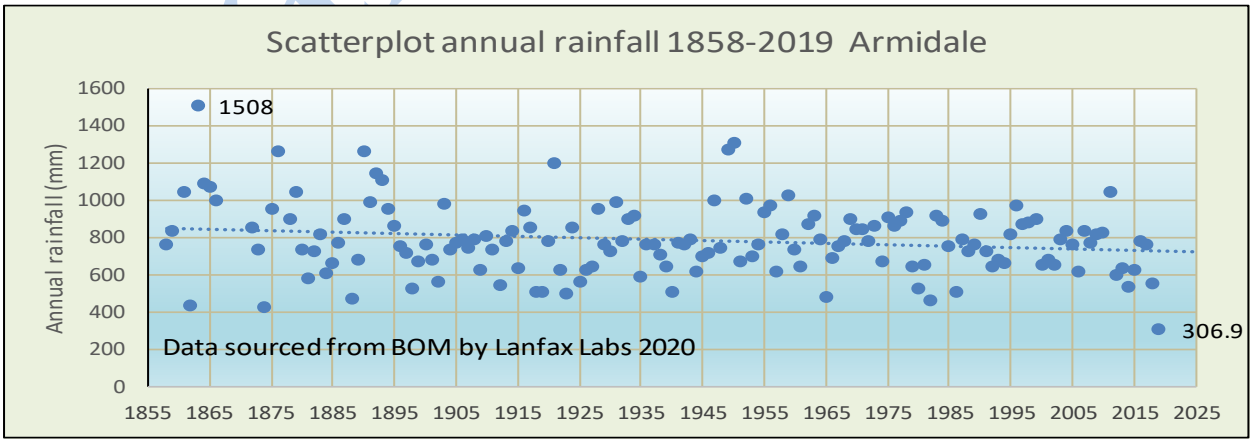

Figure 3 Scatter plot of all data for Armidale rainfall, trend line shown

The above exercise shows that you may have to refine your search for appropriate rainfall and evaporation data that more closely reflect the seasonal conditions at your wastewater site. So that 'near enough may be good enough'. The longer the period, the better. We will see how to handle the absence of long period evaporation later in this exercise.

Now we must understand some very simple statistics. When we take the annual data from the combined set above and rank the rainfall values from the highest to the lowest, the highest recorded rainfall is the 100th percentile and the lowest is the 1st percentile. The mid-range values is, of course the 50th percentile, or the 'median', that is the rainfall corresponding to the middle of the data set. Median values are not skewed by extremes on either end of the data set, simply the mid-point of the number of entries, but the mid-point does permit us to show how the data set may be skewed.

We could find the 75th percentile being 75/100 up from the lowest, or 25/100 down from the wettest. The median is simply the 50th percentile, the middle value of the whole data set.

The average is simply the sum of all the annual rainfall values divided by the number of years of data. The average need not be the same as the median. Note that in Figure 2 there are two 'outliers', data points that fall a long way off the trend line. In 1863 the annual total was 1508 mm, while in 2019 the lowest was 307 mm. As outlier, they may distort the data, but they are real recording.

Of course, there are other statistics which measure the spread of the data points from a common trend line and other measures that will not be discussed here.

3. Improvement in Statistical Choices

Once all the data have been assembled in chronological order (earliest to the latest), we need to use the statistical facilities of Excel™ to calculate a range of statistics as shown in Table 1. Here we can include columns 15 and 16 to decide which set of monthly rainfall values we will use for the water balance model.

A guide to the ideal set of monthly rainfall values can be gleaned from column 16. Compare the mean (54th percentile) with the median value of 31st percentile. Using the median would result in too small a land application area (LAA) being selected, but that is the statistic recommended for use in the NSW Guidelines (DLG *et al.*, 1998). The system designed for median rainfall would fail 70% of the time. Now look at the 90th percentile and see that it is closest to the wettest year ever, and that was shown to be an outlier in Figure 2. The LAA would be excessively large as it would never fail. If we used the highest monthly rainfall because we were 'risk adverse' the design system would be nearly three times larger than the largest year's rainfall ever recorded; simply absurd.

Table 1. A range of statistical values that may be used in a monthly water balance

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Statistic	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Monthly sum	Rank
Mean	102	87	65	45	42	55	47	48	51	67	82	88	782	779	0.54
Lowest	8	2	1	0	1	2	1	0	0	1	4	0	307	20	too dry
5th %ile	20	14	7	4	7	8	8	6	6	14	16	25	503	135	too dry
10th %ile	32	21	12	7	10	14	11	11	8	24	25	30	550	204	too dry
median	90	75	54	39	33	43	39	41	46	64	76	80	765	681	0.31
70th %ile	123	107	75	55	51	68	56	57	64	80	103	103	856	940	0.83
90th %ile	192	166	134	91	88	110	98	90	102	119	137	157	1002	1483	too wet
Highest	275	314	235	237	198	268	162	288	166	194	244	289	1508	2869	too wet
All values in millimetres Data sources www.bom.gov.au NOV20															

The mean value for the composite data in col.15 is very close to the value in col.14, simply because it is the average as there is only one way to calculate an average. However, the average rainfall occurred 54% of the whole record so the failure rate is only 41 out of every 100 (4 out of 10), very much better than using the median.

For a lower failure rate we could use a higher percentile, say 75th percentile which only fails 17 times in every 100 (twice in every 10). If we select the 90th percentile, we compare col.14 (actual recorded rainfall) with the summed monthly, we see that col. 15 is one and a half times more rain than actually recorded for the 90th percentile. It is very close to the wettest year on record in a record that stretches nearly 160 years. That does not mean that the system would not fail, because in each case the monthly value for the highest is much higher than the 90th percentile value. What we need is a percentile rank that provides a low risk of failure without huge implications for disposal area and cost. The 70-75th percentile requires a larger land application area with an acceptable risk of exceedance.

4. Calculating other percentile values

The next part of the exercise is to take the data from the Bureau's website and perform our own statistical analysis using the facilities available in spreadsheets such as Excel™.

Now return to the top right hand corner of the 'Monthly Rainfall for Armidale 56002' and find 'All years of data PDF'. Click 'All years of data' and save the zip file. Now open the folder, unzip and note three files. Open the file '……._Data12.csv as an Excel™ spreadsheet. You now have a header row and another 137 lines of yearly data set out in columns for months. You can delete column A as the information does not change. Now save the spreadsheet. You can round the numbers to integers (whole numbers) to make it easier to read, without changing the actual decimal places.

Table 2. First rows of data from the B.Met, website for Armidale

Stn	Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual
56002	1857	null	null	null	null	null	null	null	null	null	null	null	102.7	null
56002	1858	35.8	1.8	141	42.5	59.2	35.3	34.3	27.3	48	183.1	62.8	86.9	758
56002	1859	74.8	124.8	45.5	11.2	69.8	66.8	54.8	49	34.7	58.6	136.9	107.7	834.6
56002	1860	22.3	31.8	69.6	null	null	null	null	null	null	82	66.8	38.9	null
56002	1861	103.6	249.4	24.2	136.8	14.5	118	107.1	84.3	46.5	33.9	16.4	109.2	1043.9
56002	1862	56	98.3	65.4	8.2	19.3	53.9	11.2	18.8	42.8	7.1	6.9	48.3	436.2
56002	1863	274.6	164	126.4	236.5	39.8	135.4	47.6	53.1	96.3	161.6	98.3	74.4	1508
56002	1864	36	279.7	181.9	32	33.5	117.4	120.7	209	0.3	0.8	77	0.3	1088.6

Now you need to use the statistical calculating ability of Excel™ to create percentile ranks of interest. So that you can see the difference in which percentile you decide for your water balance model, I suggest that you look at median, mean, 60th, 70th and 80th percentiles.

Step 1. Create ten blank rows at the top of the sheet, copy the header into C1. Under column C insert the percentile ranks suggested above.

Step 2. In line 2 (median statistic) under the JAN column insert the formula '=MEDIAN(D$12:D$170)', where the row number block is from the 1857 year to the 2019. Don't worry about the 'null' entries as they are not counted, you don't have to change them. Be sure to fix the row number using '$', but not the column identifier.

Step 3. In line 3 (mean) under the JAN column insert the formula '=AVERAGE(D$12:D$170)', again locking the row numbers

Step 4. In line 4 (60th percentile) under the JAN column, insert the formula '=PERCENTILE.EXC(D$12:D$170,0.6)', noting that the 0.6 means the 60th percentile.

Step 5. Repeat for lines 5 and 6 for the 70th and 80th percentiles using step 4, changing the rank to 0.7 and 0.8 respectively.

Step 6. Now copy the formulae in JAN column for the five statistics into the columns E to P (annual) and there you have the statistics for each month, and the annual rainfall.

Step 7. Add two columns to the right – sum (q) and rank (r).

Into the 'sum' column for each statistic, add the formula '=SUM(D2:O2)' which sums the monthly statistics (Jan-Dec).

Into the 'rank' column, add the formula '=PERCENTRANK($P$12:$P$170,Q2)' where the row numbers correspond to the rows of data as you used above, and Q2 ranks the sum against the annual.

Copy these two cells into the four rows below for the other statistics. Now you have a set of statistical analyses for the use of these data in a water balance model as set out in Table 3.

Table 3. Data set of statistical values derived from all available rainfall data for Armidale (1857-2019)

c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r
Statistic	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Sum	rank
median	90	75	54	39	33	43	39	41	46	64	76	80	765	681	31%
mean	102	87	65	45	42	55	47	48	51	67	82	88	782	779	54%
60% percentile	106	95	67	47	42	54	47	49	54	73	91	91	791	815	62%
70th percentile	123	107	75	55	51	68	56	57	64	80	103	103	856	940	83%
90th percentile	192	166	134	91	88	110	98	90	102	119	137	157	1002	1483	98%

Which statistic you choose will be your choice, one that provides a rational sized land application area and an acceptable risk of failure. Certainly not the median value as suggested by NSW regulators which is likely to doom your client to a very high risk of failure as shown above. As noted in Table 3, the mean value has only a 30% chance of success, or in other words a 70% chance of failure; not really good odds even for a horse race.

5. Choosing the 70th percentile

There are several methods of choosing the 70th percentile monthly values. First, is that method shown in Table 2 where the calculation of the 70th percentile for the sum of the monthly values (column q) is ranked against all the actual annual rainfall (column 'p'). This method is a straight forward calculation using the spreadsheet's inbuilt capacity to provide the percentile rank.

The second method is to sort the block of monthly and annual rainfall data on the actual annual totals, from highest to lowest, then find the 70th percentile row. This method does ranks each month, as previously outlined, but shows the great variability for the years around the 70th percentile. Firstly, take the Armidale 56002 rainfall spreadsheet and sort on the annual column, highest to lowest. Since you found ranked and then sorted by rank (highest to lowest), scroll down the annual total until you find the value closest to the 70th percentile rank.

Let's consider an example. In this case, the annual 70th percentile rainfall is 867 mm. So we mark out row for the year 1976 as being closest to that value. Now pick two rows above and two rows below that year and we now have five years that represent close to the 70th percentile annual value. Table 4 shows those values for Armidale 56002/56037 data set.

Table 4. Variations in monthly rainfall around the 70th percentile annual rank

Year	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Annual	Sum	Rank
1895	256	51	40	15	23	18	13	15	56	48	127	197	859	859	0.72
1973	188	84	18	12	35	31	55	32	56	79	112	157	858	858	0.71
1917	148	112	11	3	21	19	35	35	138	69	214	51	855	855	0.70
1924	61	126	27	80	17	62	96	63	59	74	123	67	854	854	0.70
1872	203	109	80	8	5	39	56	28	53	146	48	82	854	854	0.69
Stdev	73	29	27	32	11	18	31	18	37	37	59	63	2	2	
RSD	43%	31%	77%	136%	55%	53%	61%	51%	51%	44%	48%	57%		0%	

What becomes obvious is the wide range of monthly values that reflect the variability of rainfall for a similar annual statistical outcome, as shown by the standard deviation. The coefficient of variability (CV), also known as the relative standard deviation (RSD is the ratio of the standard deviation to the mean, varying from 136% to 31%. How one could expect any of the data from Table 4 to accurately predict future rainfall event requires more than a leap of faith. However, the best models we have can only use the data from similar exercises above. The caution is that rainfall is highly variable in intensity, frequency, daily total and monthly totals. Models are simple a means of exploring the variation and from that variation make an informed 'choice'. Hard and fast rules do not apply here. Figure 4 shows the variability within this close 70th percentile annual rainfall values.

A graphical method for representing the high monthly variability is to present the data from Table 4 and Figure 4, where the monthly variation is obvious with a dominant summer rainfall and low winter rainfall. Obvious also is the significant variation for each month and no real pattern emerging.

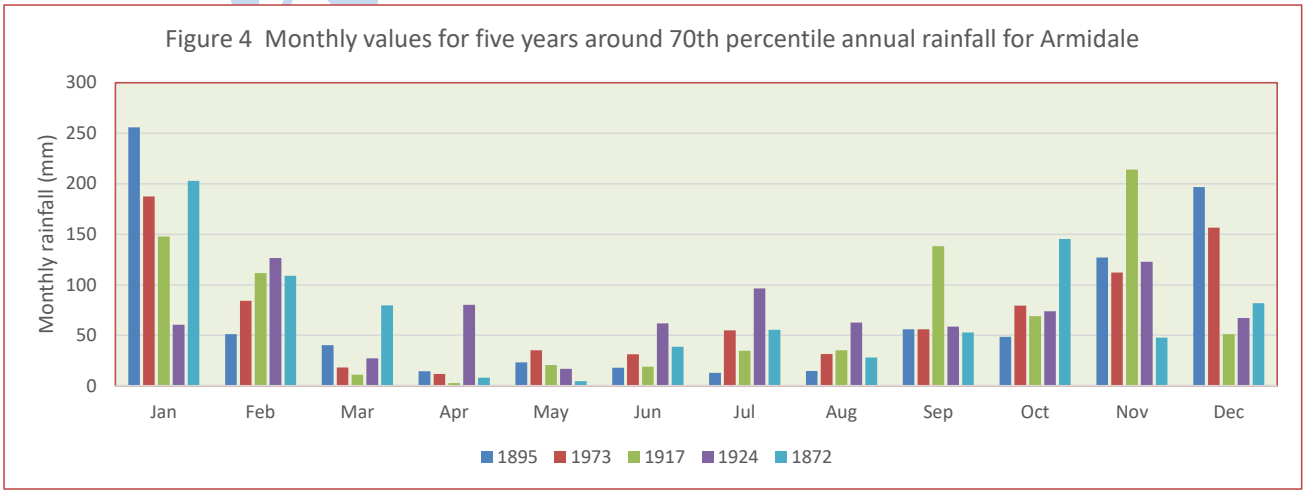

Figure 4 Graphical presentation of five years around 70th percentile

6. Modelling for all values around 70th percentile

Using a water balance model, designed along the lines of the water balance in Australian Standard AS 1547 -1994 (discontinued) and not replaced in the two subsequent versions of that Standard, the calculations for a typical soil absorption trench can be calculated. To explore the sensitivity analysis of the inputs (rainfall from Table 4) to the area of trench required, will allow us to better understand the purpose of modelling and don't expect the perfect answer.

The equation used to model the monthly outcome, that is either an area that is too large or an area that is too small, simply balances all inputs against output for each month of the year, with carry-over from one month to the next. One needs to be aware that you cannot store more water than the trench can hold, and you cannot lose by evapotranspiration and drainage from an empty trench. Think of the modelling as watching a bucket fill (water in) and empty through a hole (water out). The bucket can never be more full than full, nor more empty than empty.

Water in	= Water out	(All units in millimetres)
Rainfall + effluent	= deep drainage + evapotranspiration + change in storage in trench	

The water balance model as used in AS 1547-1994 has been set to a spreadsheet to semi-automate the calculations, a task that once configured permits easy of trying various inputs and outputs.

Water Balance Variables. For the purpose of this exercise, the porosity of the trench has been set at 70% (similar to a 250 mm high arched trench), trench width 600 mm and 450 mm deep, receiving 700 L/day of primary effluent into a soil profile of estimated LTAR of 10 L/m^2.day. A crop factor of 0.85 for October to March, and 0.6 for April to September has been set. The maximum depth of storage of water in the trench is set at the height of the tunnel (250 mm). The model is run to ensure that the trench is dry for a minimum of six months of the year.

Table 5. Outcome of modelling the five years around the 70th percentile annual rainfall.

Year of data	Estimated contact area (m^2)	Estimate max. height in trench (mm)	Estimated length of trench	No. months trench is dry
1895	68	112	62	9
1973	68	58	62	6
1917	68	95	62	6
1924	70	113	64	6
1872	69	46	63	6

The variation in estimated area of the trench is small and 70 m^2 would satisfy the water balance based upon the 70th percentile monthly rainfall.

This exercise can be easily reconfigured to suit other locations, simply by performing a statistical analysis of the monthly data. Whether there is need to look either side of the 70th percentile, or any other chosen return period will depend upon the local requirements.

Some care needs to be taken in selecting the percentile value for the modelling. From Figure 5 it is clear that the 14 annual rainfall totals that fall to the right of the central cluster are not representative of the majority of the population, and attempting to address the impact of those high rainfall years on the modelling would lead to a very long trench length. Similarly, the 10 low rainfall years may skew the statistics toward the left of average.

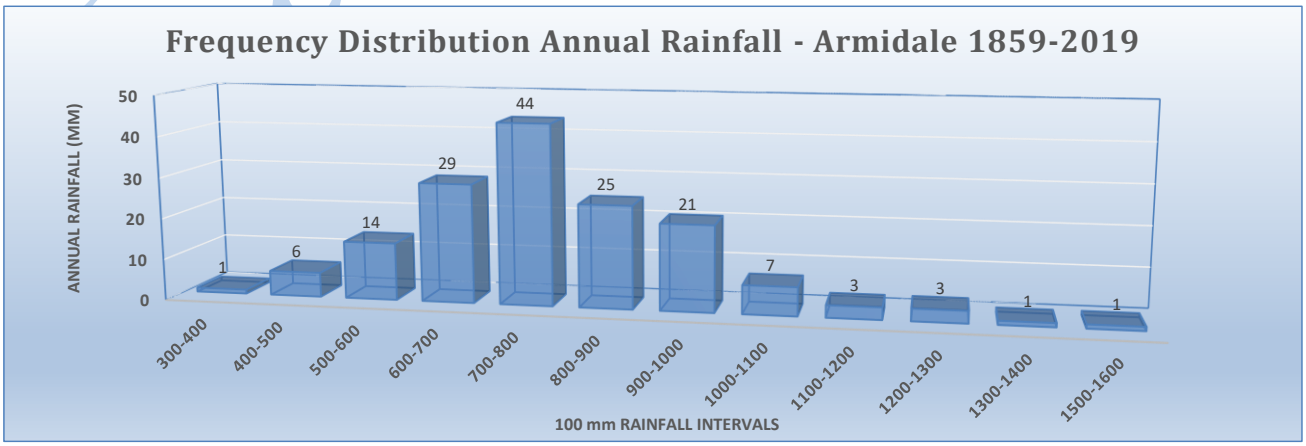

Figure 5 Distribution of whole record of rainfall for Armidale

7. Modelling historical data

Some water budgets are developed using the whole available record of all years for which historical monthly data are available, as set out n Figure 3 for Armidale. From statistical analysis of all the data, the on-site designer seeks to determine appropriately sized land application areas, based upon modelling that data, to design for an acceptable risk. It is assumed that the trench (line shown) will continue to be less influenced by the two outliers, or even remove them as anomalies and recast the data statistics. A reduce record, as shown in Figure 6 for the 142 years period omits current data since then. The trend line is similar and one outlier can be ignored.

When the modeller only uses the last short period of rainfall records, a different trend line may arise, as in Figure 7 where the data are limited to the last 23 years –for whatever reason. From that assessment one may predict that future rainfalls will be lower and a smaller land application area may satisfy.

Unfortunately, some modellers develop the water balance by averaging the number of monthly or annual failures over the 25 years. The problem with such a strategy is that one year could have 20 failures and no more for decades. How does one then decide upon the risk to the environment and human health? What are the chances of the next 25 years even slightly resembling the daily rainfall for the last 25 years?

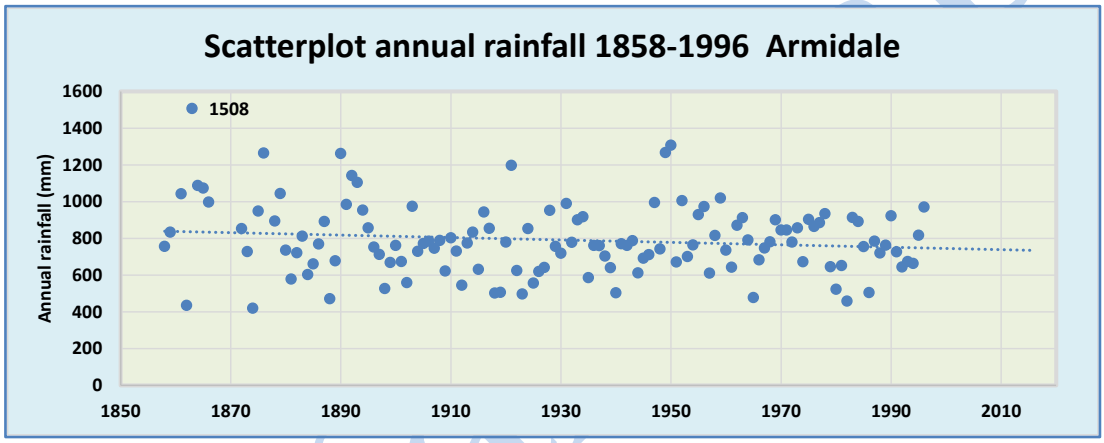

Figure 6 Limited data set 139 years Armidale, trench line shown

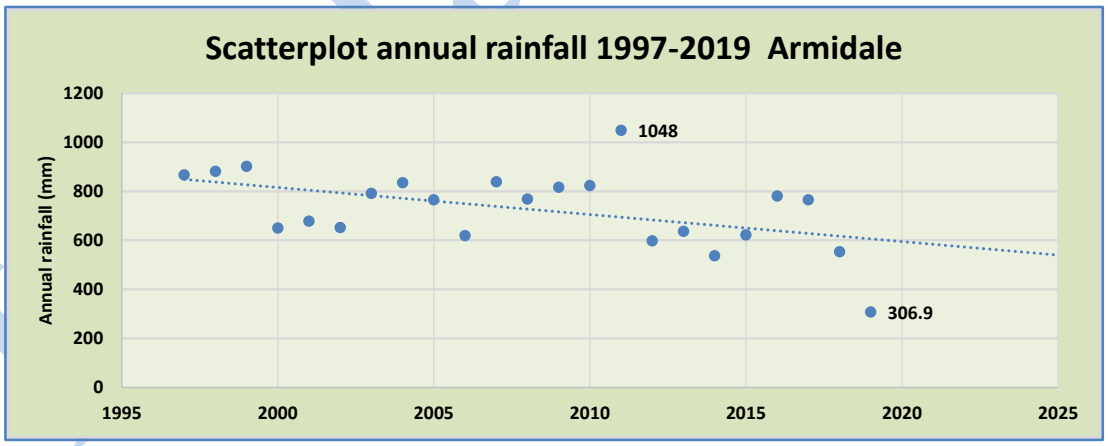

Figure 7 Limited data set 23 years to present, trench line shown

8. Conclusion

In developing a water balance model, one needs to remember that the output will only be representative of what has happened in the past. The designer needs to consider whether the outcome from the modelling using various monthly or daily historic records is likely to reflect future events. Simply adding 'fudge figures' is not good enough. Performing a series of water balance models one can glean the importance of modify the inputs that can be controlled, such as daily wastewater input and the size of the land application areas.

In this exercise, the use of the monthly rainfall from a 70th percentile year provides a reasonable estimate of the size of the trench area required.

Using slightly different parameters, similar calculations can be performed for surface or sub-surface irrigation based upon the need to limit the application rate (L/m^2.day) to avoid effluent ponding on the surface. Such modelling was beyond the scope of this paper but the principles are similar and equations are simple to modify.

9. References

Bureau of Meteorology Climate Data Online accessed from http://www.bom.gov.au/climate/data/index.shtml

DLG *et al.*, (1998) *Environment and Health Protection Guidelines On-site Sewage Management for Single Households*. Depart. Local Gov't., NSW Environ. Protection Authority, NSW Health, Land & Water Conservation, and Depart.Urban Affairs & Planning Sydney.

Standards Australia. 1994. *Australian Standard AS1547-1994 Disposal systems for effluent from domestic premises*. Standards Australia. Sydney.