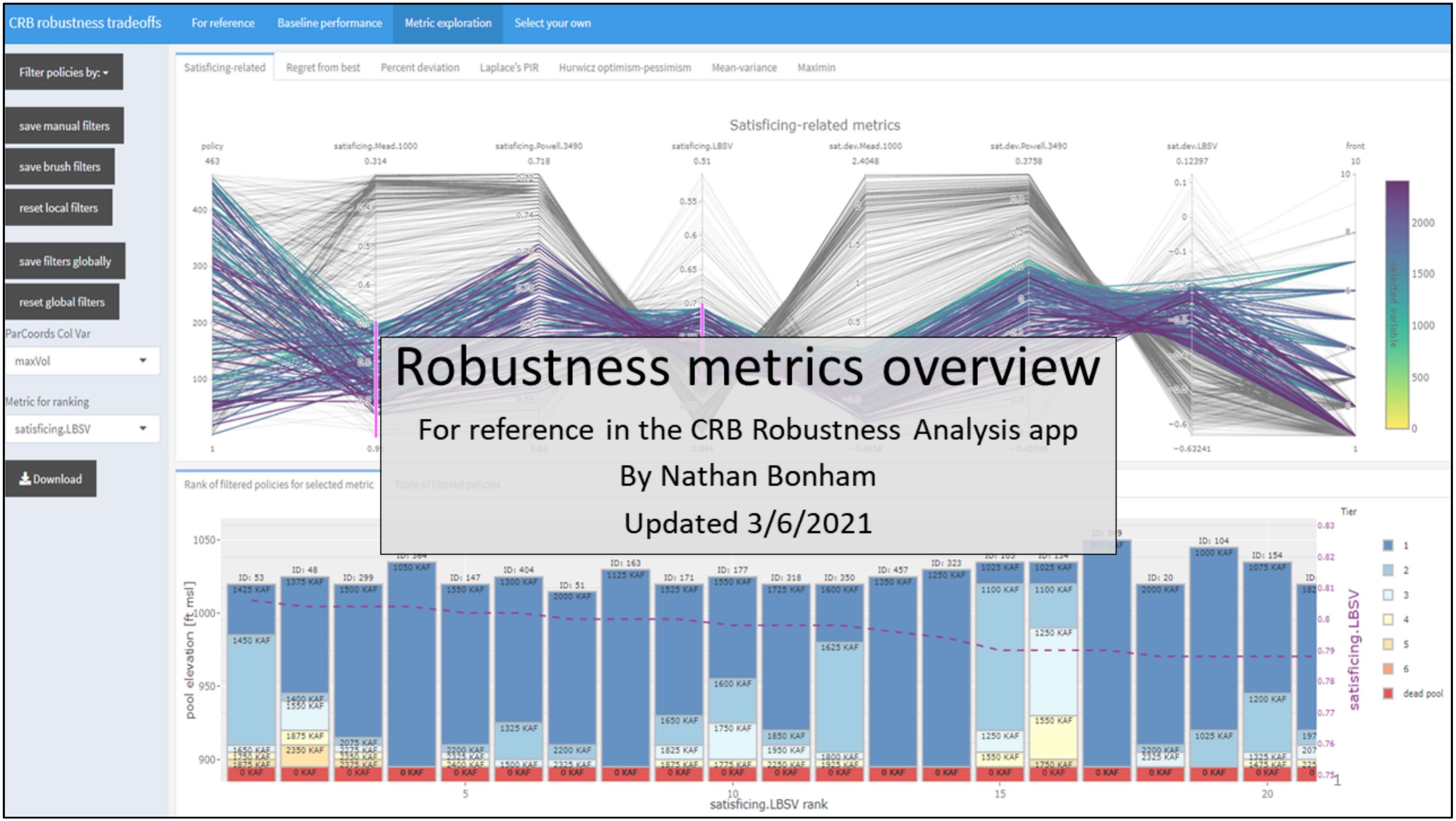
Robustness metrics overview

For reference in the CRB Robustness Analysis app

By Nathan Bonham

Updated 3/6/2021



We recommend downloading this document and having it open alongside the app for reference.

Contents

What is robustness?

- A robust solution "perform(s) well under a range of plausible conditions" (McPhail et al. 2018)
- In the Colorado River Basin (CRB) context, we desire operation policies that perform 'well' in water supply and storage objectives when tested in many plausible realizations of hydrology and demand.

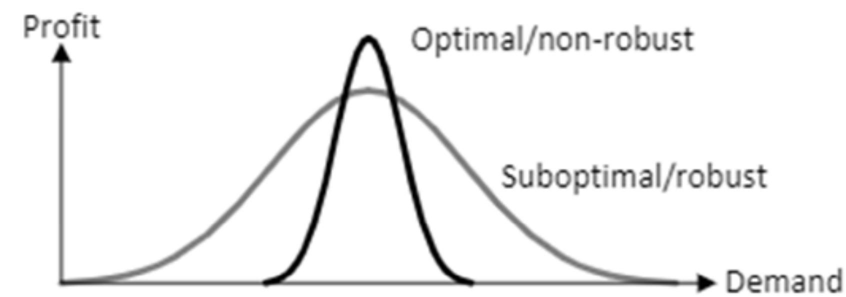


Fig. 1. Illustration of robustness.

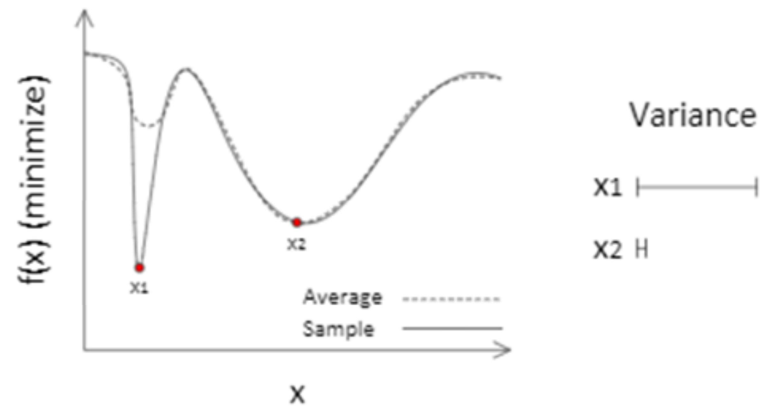
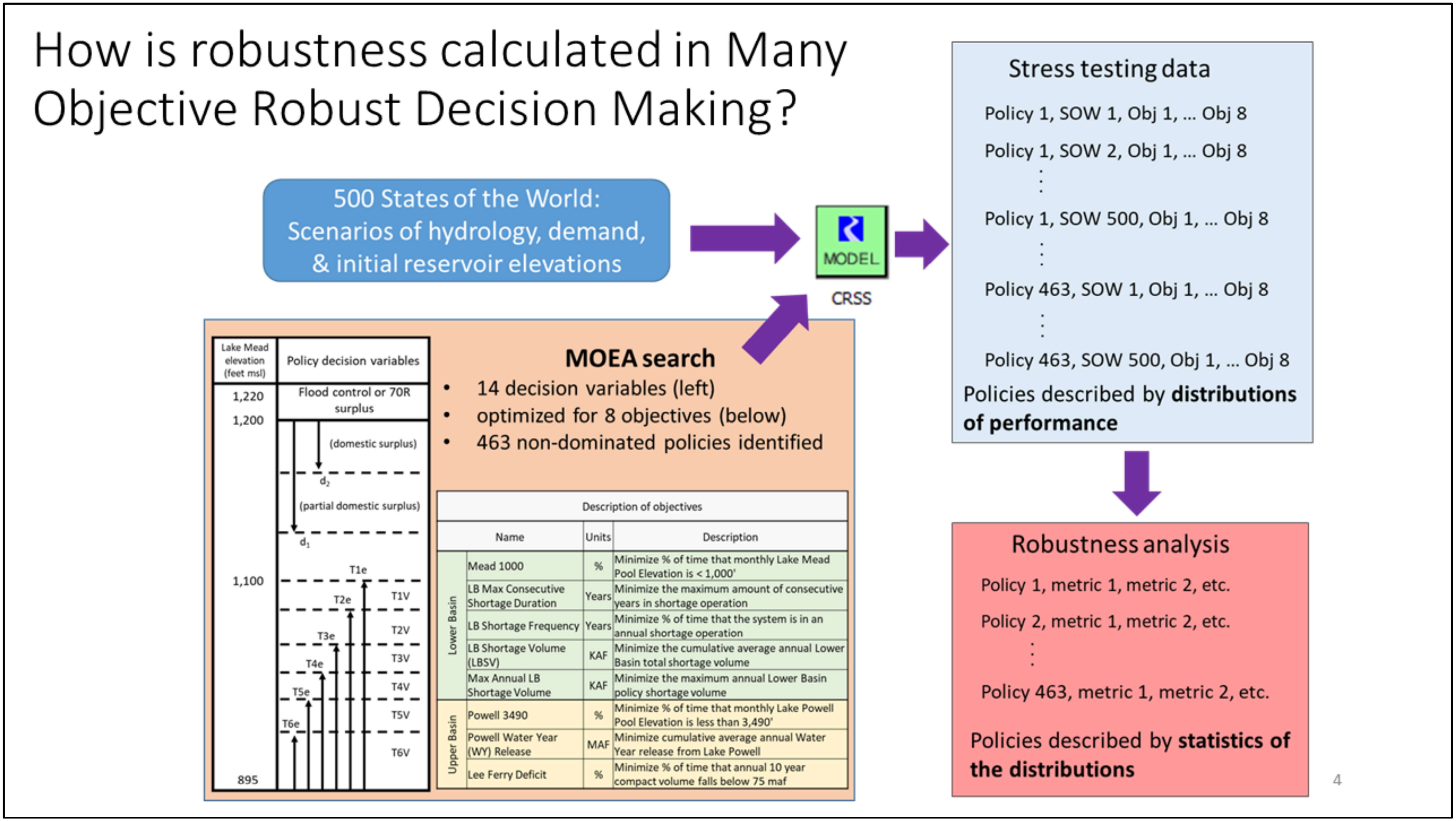


Fig. 2. Example of a sample versus a robust optimal solution for a function $f(x)$.

Figures from Syberfeldt and Gustavsson 2014

Both figures are examples where the globally optimal solution is highly sensitive to changes in another variable. In contrast, the performance of the robust solution suffers less when the uncertain variable changes.



We provide an example of how robustness is assessed in the Many Objective Robust Decision Making framework using our Colorado River Basin (CRB) case study. 463 operation policies are each simulated in 500 plausible states of the world (SOW) using the Colorado River Simulation System (CRSS). The operation policies were identified in previous research using the Borg Many Objective Evolutionary Algorithm coupled with CRSS. The result of simulating all policies in all SOW is that each policy is described by distributions of performance in the objectives. Robustness analysis is the process of summarizing the distributions of performance as scalar values by applying one or more statistical functions. The statistical functions are called robustness metrics. The figure describes the process of evaluating policy robustness, but keep in mind that copious robustness metrics exist and the analyst should consider how they are calculated and what the implications might be on robustness magnitude and ranking of the policies.

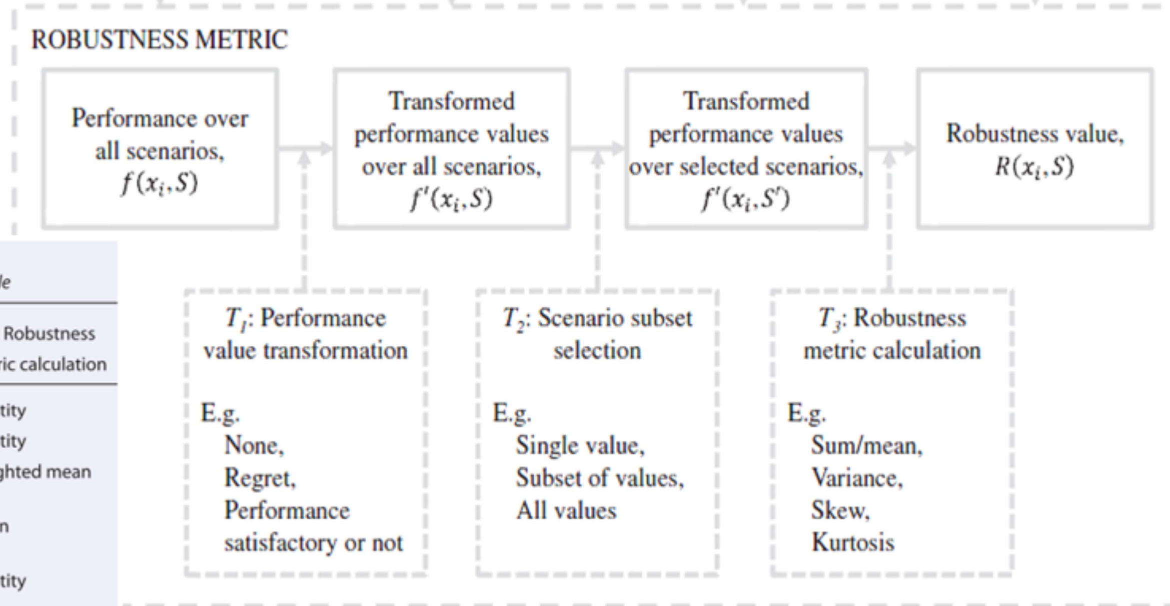
***Note: LB stands for Lower Basin

Understanding different robustness metrics

Table 1.
A Summary of the Three Transformations that are Used by Each Robustness Metric Considered in This Article

Metric	Original reference	T_1: Performance value transformation	T_2: Scenario subset selection	T_3: Robustness metric calculation
Maximin	Wald (1950)	Identity	Worst-case	Identity
Maximax	Wald (1950)	Identity	Best-case	Identity
Hurwicz optimism-pessimism rule	Hurwicz (1953)	Identity	Worst- and best-cases	Weighted mean
Laplace's principle of insufficient reason	Laplace and Simon (1951)	Identity	All	Mean
Minimax regret	Savage (1951) and Giuliani and Castelletti (2016)	Regret from best decision alternative	Worst-case	Identity
90th percentile minimax regret	Savage (1951)	Regret from best decision alternative	90th percentile	Identity
Mean-variance	Hamarat et al. (2014)	Identity	All	Mean-variance
Undesirable deviations	Kwakkel et al. (2016b)	Regret from median performance	Worst-half	Sum
Percentile-based skewness	Voudouris et al. (2014) and Kwakkel et al. (2016b)[a]	Identity	10th, 50th, and 90th percentiles	Skew
Percentile-based peakedness	Voudouris et al. (2014) and Kwakkel et al. (2016b)[a]	Identity	10th, 25th, 75th and 90th percentiles	Kurtosis
Starr's domain criterion	Starr (1963) and Schneller and Sphicas (1983)	Satisfaction of constraints	All	Mean

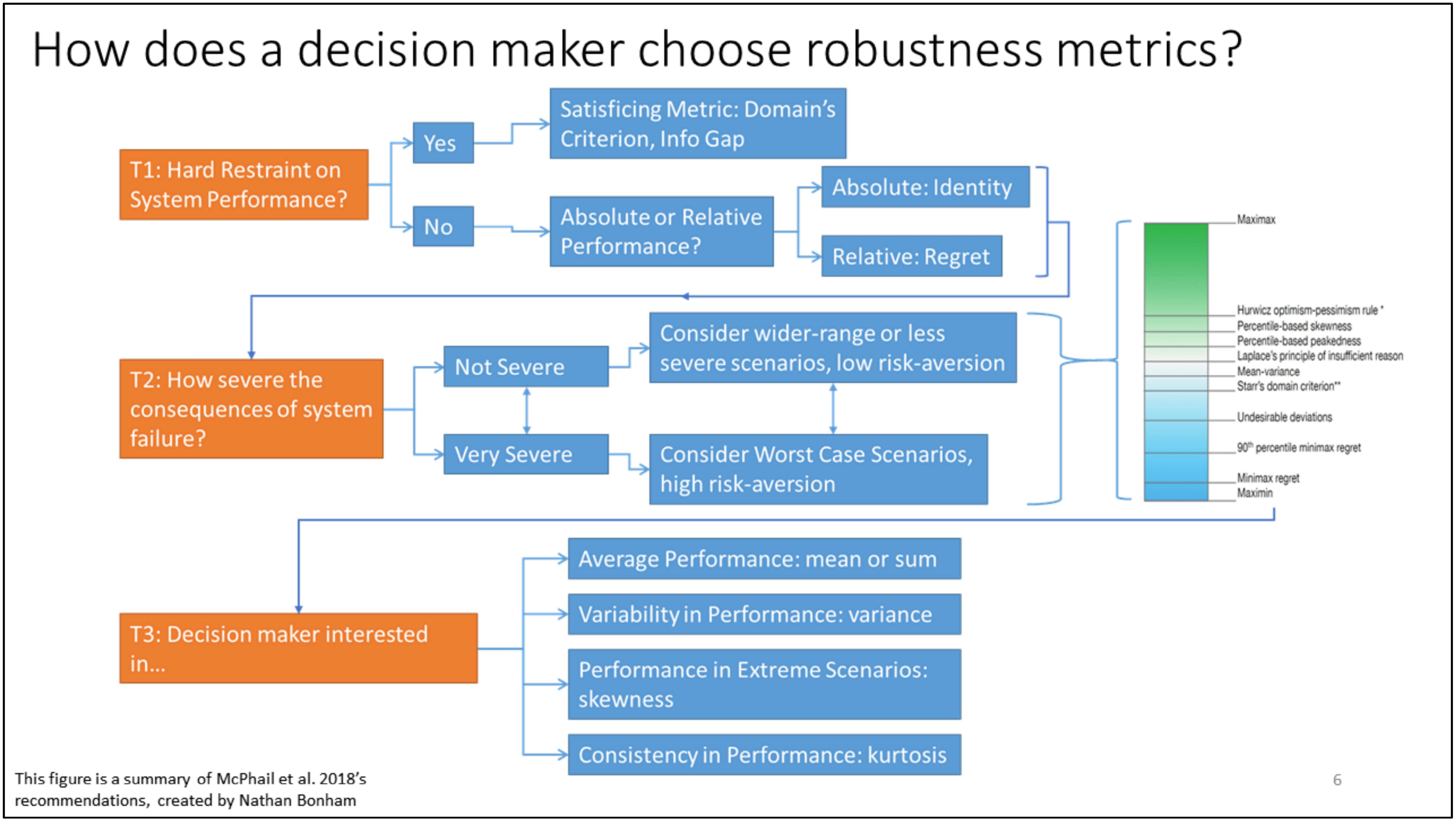
[a]Kwakkel et al. (2016b) adapted some metrics from Voudouris et al. (2014).



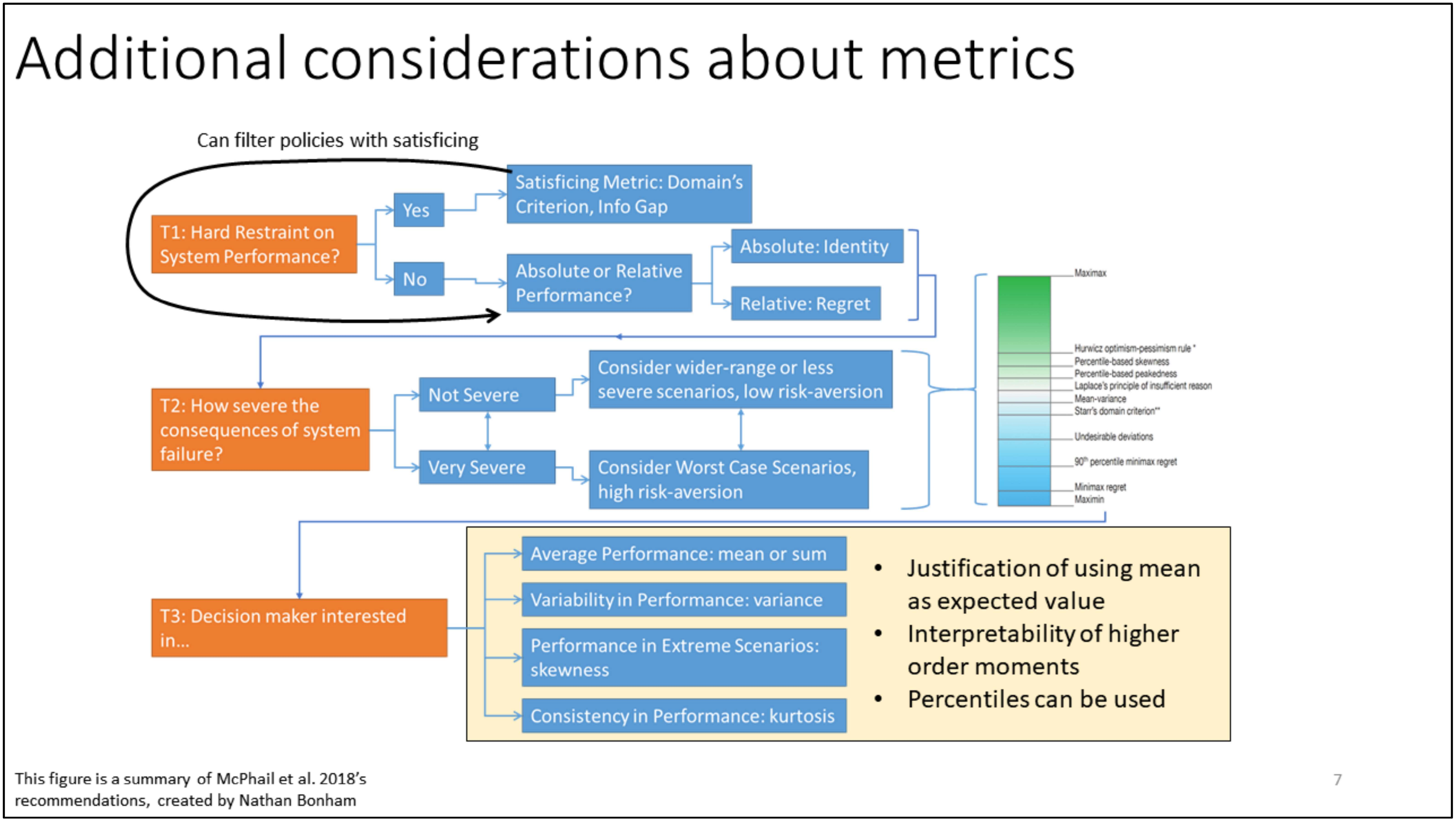
Figures from McPhail et al. 2018

5

McPhail et al. 2018 explains how robustness metrics are calculated in three steps (T1, T2, T3, right figure). T1 has to do with how the original objective values are transformed. For example, performance can be transformed to a binary (e.g., meets or fails a criteria), deviation from a benchmark (e.g., deviation from optimal performance), or left as is (e.g., Lower Basin Shortage Volume). T2 determines how many SOW are considered in the robustness metric. For example, regret from best considers only one SOW. Lastly, T3 determines how performance is summarized from a distribution of performance across multiple SOW to a single value. In other words, if the robustness metric considers more than one SOW, a statistic is used to summarize the distribution of performance as a single value. The table on the left (McPhail et al. 2018) can be referenced while using the app to understand how several common robustness metrics are calculated. Further, example calculations for metrics supported in this app are given at the end of the document.



The three steps in calculating a robustness metric can also be used to offer guidance about what robustness metrics should be used for a particular application. This figure is a summary of the suggestions from McPhail et al. 2018 created by Nathan Bonham (the green to blue gradient with labels is directly from McPhail et al. 2018). In the next slide, we add a few additional considerations to T1 and T3.



Although the diagram shows a dead-end after T1 > Yes > Satisficing Metric, satisficing metrics can be used to obtain a subset of policies that meet hard restraints, then other metrics calculated to obtain a compound definition of robustness. Further, the analyst should consider the implications of which statistic in chosen in T3. For example, taking the mean can imply that the metric represents an expected value of the future. However, this is not true when the SOW ensemble does not represent probabilistic expectations of the future. This is the case in our CRB application, where the SOW ensemble is sampling from varied hydrologic data sources and uniformly across demand and reservoir conditions. However, metrics with expected value calculations can still be used to compare policies relative to each other, but it is prudent to communicate this to stakeholders and decision makers. Further, higher-order moments are less interpretable than the mean or percentiles (e.g., average LB shortage volume has intuitive units, while kurtosis of LB shortage volume does not).

How do we handle multiple objectives in robustness metric calculations?

Objectives

	trace	policy	LBSV	Mead1000	Powell3490
1	1	1	16477.27	0.00	0.76
2	2	1	155113.64	8.33	0.00
3	3	1	0.00	0.00	0.00
4	4	1	211931.82	31.06	31.63
5	5	1	63636.36	0.00	0.00
6	6	1	300000.00	39.39	7.20
7	7	1	106818.18	8.90	0.00
8	8	1	493750.00	79.55	0.00
9	9	1	153977.27	12.50	0.00
10	10	1	164204.55	8.33	0.00

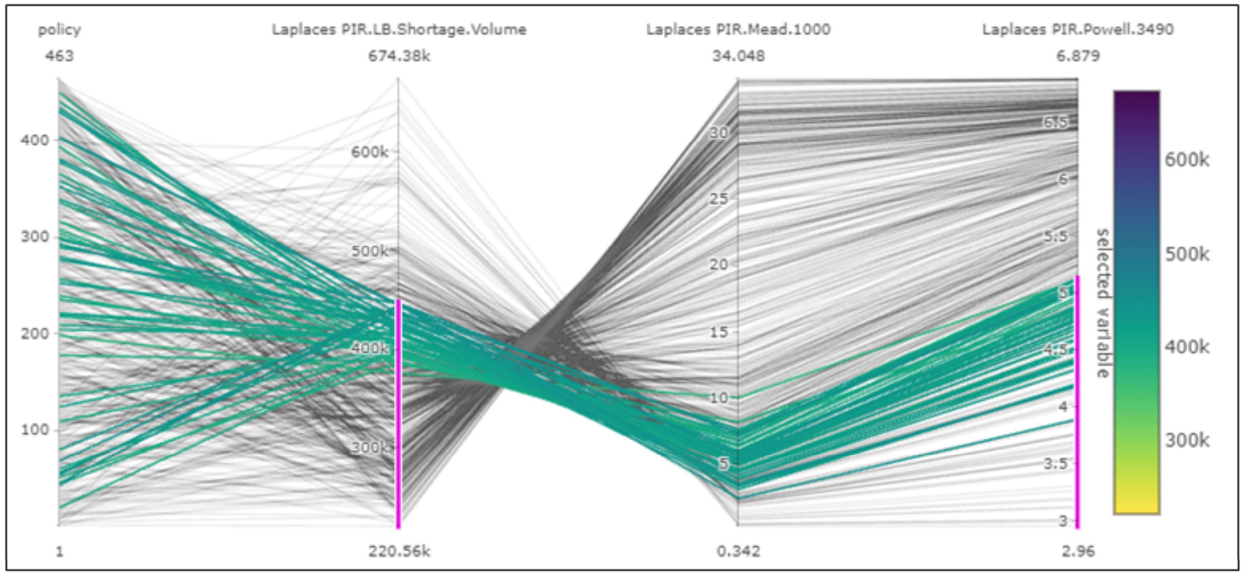
500 SOW

PIR_LBSV=mean(blue box)=216 KAF
PIR_Mead1000=mean(green box)= 26.44 %
PIR_Powell3490=mean(red box)= 6.46 %

Option 1: scale and sum

Ex: divide the mean by max, then sum
216 KAF/700 KAF + 26.44%/95% +6.46%/70% = 0.68
- pros: reducing objectives to consider
- cons: losing specificity about objective performance, have to choose scaling method, loses physical meaning (LBSV is a volume, 0.68 is a unitless score)

Option 2: explore tradeoffs interactively



In this app, we use option 2 because it is conducive to interactive exploration and explicit representation of tradeoffs.

8

When choosing robustness metrics, the analyst inherently decides how to summarize the SOW ensemble, but this does NOT generally determine how multiple objectives are handled***. Here is an example using Laplace's Principle of Insufficient Reason (PIR), which is just the average performance of the objectives across the SOW ensemble, showing two general options of how multiple objectives can be handled when quantifying robustness. In this app, we use Option 2.

*** The satisficing metric can be used to aggregate multiple objectives into one value by using the intersection or union of multiple performance thresholds as the definition of vulnerability. For example, in Alexander 2018 a SOW is classified as vulnerable if Mead 1000 < 10% or Powell 3490 < 5% or LB Shortage Volume < 600 thousand acre-feet (KAF). No scaling is used, but multiple objectives are still aggregated into one robustness metric.

Robustness metrics currently supported in this app:

- Satisficing-related
  - Satisficing fraction
  - Satisficing deviation
- Regret from best performance
- Laplace's Principle of Insufficient Reason
- Maximin
- Hurwicz optimism-pessimism rule with equal weighting
- Mean-variance
- Percent deviation from baseline

Example calculations for the supported robustness metrics

Satisficing fraction

Objective

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00
9	9	1	153977.27
10	10	1	164204.55
11	11	1	127272.73
12	12	1	13636.36
13	13	1	525000.00
14	14	1	5681.82
15	15	1	363068.18
16	16	1	407386.36
17	17	1	519318.18
18	18	1	333522.73
19	19	1	238636.36
20	20	1	101704.55

Calculations

	LBSV Bin
1	1
2	1
3	1
4	0
5	1
6	1
7	1
8	0
9	1
10	1
11	1
12	1
13	0
14	1
15	1
16	0
17	1
18	1
19	1
20	1

500 SOW

- Performance threshold: LB Policy Shortage Volume (LBSV) < 600 KAF
  - LBSV Bin = 1 means ‘threshold is met’
  - LBSV Bin = 0 mean ‘threshold is violated’
- Policy 1 example
- 500 SOW (trace column)
- Satisficing fraction = sum (LBSV Bin)/500 = 0.862
- Interpretation: policy 1 meets the 600 KAF threshold in 86.2% of the 500 SOW
- Satisficing is calculated for three performance thresholds identified in Alexander 2018:
  - Mead 1000 < 10%
  - Powell 3490 < 5%
  - LB Shortage Volume < 600 KAF
- Satisficing is NOT aggregated (this is different than Alexander 2018)

Satisficing deviation

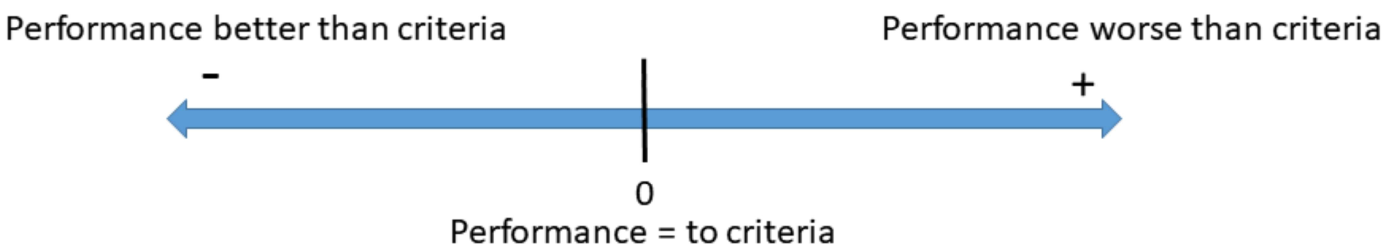
Equation 3.1 Alexander 2018, pg 39

$$\textbf{Regret Type A}\ (D_{i,j}) = \frac{|F(x_m)_{i,j} - base\ F(x_m)_i|}{base\ F(x_m)_i}$$

where x_m is an MOEA-generated operating policy; $F(x_m)_{i,j}$ is the policy's i[th] objective performance in SOW j; and $base\ F(x_m)_i$ is a policy's i[th] objective performance in the baseline SOW. Regret Type A's deviation ($\boldsymbol{D}$) is normalized by the policy's performance in the baseline SOW; thus, a robust operating policy would have minimal deviations ($\boldsymbol{D_{i,j} = 0}$) from the baseline performance across all tested SOWs.

Intuition: magnitude by which a policy is meeting satisficing criteria

$$Satisficing\ deviation\ (D_{i,j}) = \frac{F(x_m)_{i,j} - criteria\ F(x_m)_i}{criteria\ F(x_m)_i}$$



12

Satisficing deviation is similar to regret from baseline (regret type A) from Alexander 2018. First, the baseline performance is defined as the satisficing thresholds instead of performance in MOEA optimization. Second, absolute value is not taken in the numerator.

Satisficing deviation: calculation example

Objectives

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00
9	9	1	153977.27
10	10	1	164204.55

500 SOW

Calculations

	LBSV
1	-0.97253788
2	-0.74147727
3	-1.00000000
4	-0.64678030
5	-0.89393940
6	-0.50000000
7	-0.82196970
8	-0.17708333
9	-0.74337122
10	-0.72632575

- Performance threshold: LB Policy Shortage Volume (LBSV) < 600 KAF
- Policy 1 example
- 500 SOW

- Example calculation in trace (SOW) 1:
  - (16,477.27 AF – 600,000 AF)/ 600,000 AF = -0.973
- SOW are aggregated by taking the mean

Regret from best performance

Equation 3.2

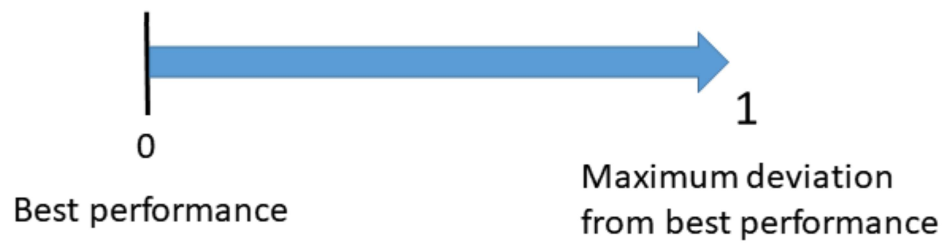
Alexander 2018, pg 40

$$\textbf{Regret Type B } (D_{i,j}) = \frac{|F(x_m)_{i,j} - best\ F(x_m)_{i,j}|}{F(x_m)_{i,j}}$$

where $best\ F(x_m)_{i,j}$ is the best value of the i[th] objective in SOW j. Regret Type B's deviation (**D**) is normalized by the objective values since the best value often approaches zero in minimization optimizations formulations. Therefore, a robust operating policy for Regret Type B would have low deviations from all the best objective performance values across all tested SOWs.

- Example: SOW S, policy A vs policy B
  - A: Mead 1000 = 20%
  - B: Mead 1000 = 5%
  - Best = 0 %
  - Worst = 40 %
- Regret with original denominator in Alexander 2018:
  - A: (20 – 0)/20 = 1
  - B: (5-0)/5 = 1
  - Policy and A and B have same normalized value!
- Regret with maximum deviation as denominator
  - A: (20 – 0)/(40-0) = 0.5
  - B: (5 – 0)/(40-0) = 0.125
  - A has larger regret than B

$$Regret\ Type\ B\ (D_{i,j}) = \frac{|F(x_m)_{i,j}\ - best\ F(x_m)_{i,j}|}{|best\ F(x_m)_{i,j} - worst\ F(x_m)_{i,j}|}$$



14

Regret from best performance is regret type B from Alexander 2018 with a different normalization factor. The denominator is the maximum deviation from best performance of all policies in objective if for SOW j. We chose this as the normalization factor is consistent for all policies, which avoids the case where different normalization factors can result in the same regret despite different performance values. This slide shows example calculations for the original normalization factor and the new normalization factor, plus a figure showing interpretation.

Regret from best performance example calculation

Objectives

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00

500 SOW

Calculations

	LBSV
1	0.155913950
2	0.366935495
3	0.000000000
4	0.059629598
5	0.225806440
6	0.323954685
7	0.353198690
8	0.393675156

Best performance

	LB.Shortage.Volume.Policy
1	0.00
2	0.00
3	0.00
4	152840.91
5	0.00
6	43181.82
7	0.00
8	22727.27
9	2840.91
10	0.00

Worst performance

	LB.Shortage.Volume.Policy
1	105681.82
2	422727.27
3	0.00
4	1143807.01
5	281818.18
6	835941.40
7	302430.85
8	1219202.90
9	456903.45
10	457954.55

500 SOW

*Best and worst performance are global across policies

Policy 1, SOW 4 example :

- LBSV: $|\underbrace{211{,}931}_{\text{actual}} - \underbrace{152{,}840}_{\text{best}}| \;/\; |\underbrace{152{,}840}_{\text{best}} - \underbrace{1{,}143{,}807}_{\text{worst}}| = 0.0596$
- SOW aggregation: mean

15

Laplace's Principle of Insufficient Reason (PIR)

Objectives

	trace	policy	LBSV	Mead1000	Powell3490
1	1	1	16477.27	0.00	0.76
2	2	1	155113.64	8.33	0.00
3	3	1	0.00	0.00	0.00
4	4	1	211931.82	31.06	31.63
5	5	1	63636.36	0.00	0.00
6	6	1	300000.00	39.39	7.20
7	7	1	106818.18	8.90	0.00
8	8	1	493750.00	79.55	0.00
9	9	1	153977.27	12.50	0.00
10	10	1	164204.55	8.33	0.00

500 SOW

- Value of each objective averaged over the SOW ensemble
- Example:
  - PIR_LBSV=mean(blue box)=216 KAF
  - PIR_Mead1000=mean(green box)= 26.44 %
  - PIR_Powell3490=mean(red box)= 6.46 %

Maximin

Objectives

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00
9	9	1	153977.27
10	10	1	164204.55

500 SOW

Best: 0 KAF
Worst: 575 KAF

- Maximin: value of worst case SOW
  - High-risk aversion
- example with Lower Basin Shortage Volume
- Maximin = 575 KAF

Hurwicz optimism-pessimism rule

Objectives

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00
9	9	1	153977.27
10	10	1	164204.55

500 SOW

Best: 0 KAF
Worst: 575 KAF

- Weighted average of best and worst case SOW
- This app uses equal weighting for best and worst case SOW (0.5 each)
- Example:
  - Best weight = 0.5
  - Worst weight= 1-0.5=0.5
  - Result = (0.5*0 + 575 KAF *0.5)/2 = 143.75 KAF

18

Mean-variance

Objectives

	trace	policy	LBSV
1	1	1	16477.27
2	2	1	155113.64
3	3	1	0.00
4	4	1	211931.82
5	5	1	63636.36
6	6	1	300000.00
7	7	1	106818.18
8	8	1	493750.00
9	9	1	153977.27
10	10	1	164204.55

500 SOW

mean: 216 KAF
Standard deviation: 165 KAF

- In case of maximization: mean/standard deviation
- In case of minimization: mean * standard deviation (all objectives in this app are minimization)
- Example with Lower Basin Shortage Volume
  - LBSV is minimized
  - Use mean * standard deviation
- Result: 216 KAF * 165 KAF = 35 Billion-AF squared
- Interpretation: want to minimize the metric. The metric is penalized if either the mean or standard deviation increases

19

Percent deviation from baseline

TraceNumber	policy	Mead.1000	CDF
285	1	98.86	1.000
63	1	97.92	0.998
324	1	96.40	0.996
171	1	94.89	0.994
86	1	93.37	0.992
388	1	92.42	0.990
313	1	92.23	0.988
264	1	72.54	0.908
419	1	72.54	0.908
199	1	72.35	0.904
355	1	72.35	0.904
387	1	71.78	0.900

500 SOW

Kazprzyk et al. 2013

$$d_i = \begin{cases} \dfrac{f_{i\ 90} - f_{i\ \text{base}}}{f_{i\ \text{base}}} & \text{if } i \text{ is minimized} \\ \dfrac{f_{i\ \text{base}} - f_{i\ 10}}{f_{i\ \text{base}}} & \text{if } i \text{ is maximized} \end{cases}$$

i: performance measure of interest
90[10]: percentile in SOW ensemble
Base: performance in optimization

90th percentile performance: 71.8%
Baseline performance: 25.98 %

Deviation = (71.8% - 25.98%)/25.98%= 1.765 = 177%

20

If baseline performance is zero (ie, denominator is zero), then the code implements an if statement. If the numerator is also zero (or within some very small user-defined tolerance), then percent deviation is set to zero. If not, then percent deviation is set to 999 to indicate that base performance is zero and deviation is greater than zero.

References

- Alexander, E., 2018. Searching for A Robust Operation of Lake Mead (M.S.). University of Colorado, Boulder, CO.
- Kasprzyk, J.R., Nataraj, S., Reed, P.M., Lempert, R.J., 2013. Many objective robust decision making for complex environmental systems undergoing change. Environ. Model. Softw. 42, 55–71. https://doi.org/10.1016/j.envsoft.2012.12.007
- McPhail, C., Maier, H.R., Kwakkel, J.H., Giuliani, M., Castelletti, A., Westra, S., 2018. Robustness Metrics: How Are They Calculated, When Should They Be Used and Why Do They Give Different Results? Earths Future 6, 169–191. https://doi.org/10.1002/2017EF000649
- Syberfeldt, A., Gustavsson, P., 2014. Increased Robustness of Product Sequencing Using Multi-objective Optimization. Procedia CIRP 17, 434–439. https://doi.org/10.1016/j.procir.2014.01.141

Please email me with questions and comments

Nathan.Bonham@Colorado.edu