Statistical Performances measures - models comparison

L Patryl\textsuperscript{a}, D. Galeriu\textsuperscript{a} ...

\textsuperscript{a} Commissariat à l’Energie Atomique, DAM, DIF, F-91297 Arpajon (France)
\textsuperscript{b} "Horia Hulubei" Institute for Physics & Nuclear Engineering (Romania)

September, 12\textsuperscript{th} 2011
OUTLINE

1. Statistical performance measure

2. Simple statistical analysis on wheat experiments

3. Conclusions
1. Statistical performance measure

2. Simple statistical analysis on wheat experiments

3. Conclusions
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
**Introduction.**

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)
Introduction.

In order to compare predictions from a model and observations measurements, several statistical performances measures can be used (U.S. Environmental Protection Agency).

Some of these performance measures are:

- the fractional bias (FB)
- the geometric mean bias (MG);
- the normalized mean square error (NMSE);
- the geometric variance (VG)
- the correlation coefficient (R)
- the fraction of predictions within a factor of two of observations (FAC2)

A perfect model would have

\[
\text{MG, VG, R, and FAC2}=1.0;
\]

\[
\text{FB and NMSE} = 0.0.
\]
Systematic errors

- the systematic bias refers to the ratio of \( C_p \) to \( C_o \).
- FB and MG are measures of mean bias and indicate only systematic errors which lead to always underestimate or overestimate the measured values.
- FB is based on a linear scale and the systematic bias refers to the arithmetic difference between \( C_p \) and \( C_o \).
- MG is based on a logarithmic scale.
Systematic errors

- The systematic bias refers to the ratio of \( C_p \) to \( C_o \).
- FB and MG are measures of mean bias and indicate only systematic errors which lead to always underestimate or overestimate the measured values.

FB is based on a linear scale and the systematic bias refers to the arithmetic difference between \( C_p \) and \( C_o \).

MG is based on a logarithmic scale.

\[
FB = \frac{\sum_i (C_{oi} - C_{pi})}{0.5 \sum_i (C_{oi} + C_{pi})} = FB_{FN} - FB_{FP}
\]
Systematic errors.

- the systematic bias refers to the ratio of $C_p$ to $C_o$
- $FB$ and $MG$ are measures of mean bias and indicate only systematic errors which lead to always underestimate or overestimate the measured values,
- $FB$ is based on a linear scale and the systematic bias refers to the arithmetic difference between $C_p$ and $C_o$,
- $MG$ is based on a logarithmic scale.

\[
FB = \frac{\sum_i (C_{oi} - C_{pi})}{0.5 \sum_i (C_{oi} + C_{pi})} = FB_{FN} - FB_{FP}
\]
Systematic errors.

- The systematic bias refers to the ratio of Cp to Co.
- FB and MG are measures of mean bias and indicate only systematic errors which lead to always underestimate or overestimate the measured values.
- FB is based on a linear scale and the systematic bias refers to the arithmetic difference between Cp and Co.
- MG is based on a logarithmic scale.

\[ MG = e^{(\ln C_o - \ln C_p)} \]
Random errors

Systematic and Random errors.

- Random error is due to unpredictable fluctuations We don’t have expected value
- Values are scattered about the true value, and tend to have null arithmetic mean when measurement is repeated.
- NMSE and VG are measures of scatter and reflect both systematic and unsystematic (random) errors.
Random errors

Systematic and Random errors.

- Random error is due to unpredictable fluctuations. We don’t have expected value.
- Values are scattered about the true value, and tend to have null arithmetic mean when measurement is repeated.
- NMSE and VG are measures of scatter and reflect both systematic and unsystematic (random) errors.
Random errors

Systematic and Random errors.

- Random error is due to unpredictable fluctuations. We don’t have expected value.
- Values are scattered about the true value, and tend to have null arithmetic mean when measurement is repeated.
- NMSE and VG are measures of scatter and reflect both systematic and unsystematic (random) errors.

\[
NMSE = \frac{(C_o - C_P)^2}{(C_o C_P)}
\]
Random errors

Systematic and Random errors.

- Random error is due to unpredictable fluctuations. We don’t have expected value.
- Values are scattered about the true value, and tend to have null arithmetic mean when measurement is repeated.
- NMSE and VG are measures of scatter and reflect both systematic and unsystematic (random) errors.

\[ VG = e^{(\ln C_o - \ln C_p)} \]
Reflects the linear relationship between two variables

- It is insensitive to either an additive or a multiplicative factor
- A perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model.
- For example, a scatter plot might show generally poor agreement, however, the presence of a good match for a few extreme pairs will greatly improve $R$.

To avoid using

$$ R = \frac{(C_o - \overline{C}_0)(C_p - \overline{C}_p)}{\sigma_{C_o} \sigma_{C_p}} $$
Correlation coefficient $R$

- Reflects the linear relationship between two variables
- It is insensitive to either an additive or a multiplicative factor

A perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model. For example, scatter plot might show generally poor agreement, however, the presence of a good match for a few extreme pairs will greatly improve $R$.

To avoid using

$$ R = \frac{(C_o - \overline{C_0}) (C_p - \overline{C_p})}{\sigma_{C_o} \sigma_{C_p}} $$
Correlation coefficient $R$.

- Reflects the linear relationship between two variables
- It is insensitive to either an additive or a multiplicative factor
- A perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model.

For example, a scatter plot might show generally poor agreement, however, the presence of a good match for a few extreme pairs will greatly improve $R$.

to avoid using

$$R = \frac{(C_o - \overline{C_0})(C_p - \overline{C_p})}{\sigma_{C_0} \sigma_{C_p}}$$
Correlation coefficient $R$

- Reflects the linear relationship between two variables
- It is insensitive to either an additive or a multiplicative factor
- A perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model.
- For example, scatter plot might show generally poor agreement, however, the presence of a good match for a few extreme pairs will greatly improve $R$.

$$R = \frac{\left( C_o - \bar{C}_0 \right) \left( C_p - \bar{C}_p \right)}{\sigma_{C_o} \sigma_{C_p}}$$
Correlation coefficient R.

- Reflects the linear relationship between two variables
- It is insensitive to either an additive or a multiplicative factor
- A perfect correlation coefficient is only a necessary, but not sufficient, condition for a perfect model.
- For example, a scatter plot might show generally poor agreement, however, the presence of a good match for a few extreme pairs will greatly improve R.
- **to avoid using**

\[
R = \frac{(C_o - \overline{C_0})(C_p - \overline{C_p})}{\sigma_{C_o} \sigma_{C_p}}
\]
FAC2

FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.

\[ \text{FAC2} = \text{fraction of data that satisfy } 0.5 \leq \frac{C_p}{C_o} \leq 2.0 \]
### Properties of Performance measures.

- **Multiple performance measures have to be considered**
  - Advantages of each performance measure are partly determined by the distribution of the variable.
  - For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values.
  - MG and VG would be more appropriate for a dataset were both predicted and observed concentrations vary by many orders of magnitude.
  - However, MG and VG are strongly influenced by extremely low values and are undefined for zero values. It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD.
  - FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
  - FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.
Properties of Performance measures.

- multiple performance measures have to be considered
- Advantages of each performance measure are partly determined by the distribution of the variable
  - For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values
  - MG and VG would be more appropriate for a dataset where both predicted and observed concentrations vary by many orders of magnitude.
  - However, MG and VG are strongly influenced by extremely low values and are undefined for zero values.
  - It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD).
  - In this case, if Cp or Co are lower than the threshold, they are set to the LOD.
  - FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
  - FAC2 is the most robust measure, because it is not overly influenced by high and low outliers.
Properties of Performance measures.

- multiple performance measures have to be considered.
- Advantages of each performance measure are partly determined by the distribution of the variable.
- For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values.
- MG and VG would be more appropriate for dataset were both predicted and observed concentrations vary by many orders of magnitude.
- However, MG and VG are strongly influenced by extremely low value and are undefined for zero values → It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD.
- FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
- FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.
Properties of Performance measures.

- multiple performance measures have to be considered
- Advantages of each performance measure are partly determined by the distribution of the variable
- For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values
- MG and VG would be more appropriate for dataset were both predicted and observed concentrations vary by many orders of magnitude.
  - However, MG and VG are strongly influenced by extremely low value and are undefined for zero values → It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD
- FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
- FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.
Properties of Performance measures

- multiple performance measures have to be considered
- Advantages of each performance measure are partly determined by the distribution of the variable
- For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values
- MG and VG would be more appropriate for a dataset where both predicted and observed concentrations vary by many orders of magnitude.
- However, MG and VG are strongly influenced by extremely low values and are undefined for zero values → It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD
- FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
- FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.
Properties of Performance measures

- Multiple performance measures have to be considered.
- Advantages of each performance measure are partly determined by the distribution of the variable.
- For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values.
- MG and VG would be more appropriate for datasets where both predicted and observed concentrations vary by many orders of magnitude.
- However, MG and VG are strongly influenced by extremely low values and are undefined for zero values. → It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD.
- FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentrations.
- FAC2 is the most robust measure, because it is not overly influenced by high and low outliers.
Properties of Performance measures.

- multiple performance measures have to be considered
- Advantages of each performance measure are partly determined by the distribution of the variable
- For a log normal distribution, MG and Vg provide a more balanced treatment of extremely high and low values
- MG and VG would be more appropriate for dataset were both predicted and observed concentrations vary by many orders of magnitude.
- However, MG and VG are strongly influenced by extremely low value and are undefined for zero values → It is necessary to impose a minimum threshold for data which can be the limit of detection (LOD). In this case, if Cp or Co are lower than the threshold, they are set to the LOD
- FB and NMSE are strongly influenced by infrequently occurring high observed and predicted concentration.
- FAC2 is the most robust measure, because it is not overly influenced by high and low outlier.
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{C_p}{C_o} = \frac{1-0.5FB}{1+0.5FB}
\]
Interpretation of Performance measures

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction)
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias
- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction

\[
\frac{C_p}{C_o} = \frac{1-0.5FB}{1+0.5FB}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).

- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.

- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.

- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.

- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.

- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.

- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.

- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.

- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.

- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.

- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{C_p}{C_o} = \frac{1 - 0.5FB}{1 + 0.5FB}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Values of NMSE that are equal to 0.5 correspond to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Values of VG that are equal to 1.6 correspond to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{C_p}{C_o} = \frac{1-0.5FB}{1+0.5FB}
\]
Interpretation of Performance measures

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{C_p}{C_o} = \frac{1 - 0.5FB}{1 + 0.5FB}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Values of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Values of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{\langle C_p \rangle}{\langle C_0 \rangle} = \frac{1}{MG}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{\langle C_p \rangle}{\langle C_0 \rangle} = \frac{1}{MG}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn't differentiate whether the factor of two mean bias is underprediction or overprediction.

\[ \frac{C_P}{C_O} = \frac{2 + \text{NMSE} \pm \sqrt{(2 + \text{NMSE})^2 - 4}}{2} \]
Interpretation of Performance measures

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction)
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction

\[
\frac{C_p}{C_0} = \frac{2 + NMSE \pm \sqrt{(2 + NMSE)^2 - 4}}{2}
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{\langle C_p \rangle}{\langle C_o \rangle} = \exp[\pm \sqrt{lnVG}]
\]
Interpretation of Performance measures.

- FB is symmetrical and bounded; values for the fractional bias range between -2.0 (extreme underprediction) to +2.0 (extreme overprediction).
- The fractional bias is a dimensionless number, which is convenient for comparing the results from studies involving different concentration levels.
- Values of the FB that are equal to -0.67 are equivalent to underprediction by a factor of two.
- Values of the FB that are equal to +0.67 are equivalent to overprediction by a factor of two.
- Model predictions with a fractional bias of 0 (zero) are relatively free from bias.
- Values of the MG that are equal to +0.5 are equivalent to underprediction by a factor of two.
- Values of the MG that are equal to +2 are equivalent to overprediction by a factor of two.
- Value of NMSE that are equal to 0.5 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.
- Value of VG that are equal to 1.6 corresponds to an equivalent factor of two mean bias.
- It doesn’t differentiate whether the factor of two mean bias is underprediction or overprediction.

\[
\frac{\langle C_p \rangle}{\langle C_0 \rangle} = \exp[\pm \sqrt{\ln VG}]
\]
Model acceptance Criteria

How good is good enough?

- Fraction of prediction within a factor 2 of observation is about 50% or greater ($FAC2 > 0.5$)
- The mean bias is within $\pm 30\%$ of the mean ($|FB| < 0.3$ or $0.7 < MG < 1.3$)
- Random scatter is about a factor of two to three of the mean ($NMSE < 1.5$ or $VG < 4$)
Model acceptance Criteria

How good is good enough?

- Fraction of prediction within a factor 2 of observation is about 50% or greater ($FAC2 > 0.5$)
- The mean bias is within ±30% of the mean ($|FB| < 0.3$ or $0.7 < MG < 1.3$)
- Random scatter is about a factor of two to three of the mean ($NMSE < 1.5$ or $VG < 4$)
How good is good enough?

- Fraction of prediction within a factor 2 of observation is about 50% or greater ($FAC2 > 0.5$)
- The mean bias is within ±30% of the mean ($|FB| < 0.3$ or $0.7 < MG < 1.3$)
- Random scatter is about a factor of two to three of the mean ($NMSE < 1.5$ or $VG < 4$)
OUTLINE

1 Statistical performance measure

2 Simple statistical analysis on wheat experiments

3 Conclusions
Difficult to say which model is better

Difficult to say if models make overprediction or underprediction
Difficult to say which model is better

Difficult to say if models make overprediction or underprediction

Predicted (Bq l⁻¹)

Measured (Bq l⁻¹)
61 experiments

- 3 models (CEA, JAEA, IFIN)
- Some of values equal 0 → without detection threshold or other informations we use only arithmetic scale (FB and NMSE)
- More than a factor 2 for CEA and JAEA (random and systematic errors)
- Only about 30% value are within a factor of 2 of observations

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.16</td>
<td>0.31</td>
<td>0.858</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.13</td>
<td>0.26</td>
<td>0.30</td>
<td>0.818</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.42</td>
<td>0.15</td>
<td>0.36</td>
<td>0.912</td>
</tr>
</tbody>
</table>
61 experiments

3 models (CEA, JAEA, IFIN)

Some of values equal 0 without detection threshold or other informations we use only arithmetic scale (FB and NMSE)

More than a factor 2 for CEA and JAEA (random and systematic errors)

Only about 30% values are within a factor of 2 of observations

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.16</td>
<td>0.31</td>
<td>0.858</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.13</td>
<td>0.26</td>
<td>0.30</td>
<td>0.818</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.42</td>
<td>0.15</td>
<td>0.36</td>
<td>0.912</td>
</tr>
</tbody>
</table>
61 experiments
3 models (CEA, JAEA, IFIN)

- Some of values equal 0 → without detection threshold or other informations we use only arithmetic scale (FB and NMSE)
- More than a factor 2 for CEA and JAEA (random and systematic errors)
- Only about 30% value are within a factor of 2 of observations

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.16</td>
<td>0.31</td>
<td>0.858</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.13</td>
<td>0.26</td>
<td>0.30</td>
<td>0.818</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.42</td>
<td>0.15</td>
<td>0.36</td>
<td>0.912</td>
</tr>
</tbody>
</table>
61 experiments
3 models (CEA, JAEA, IFIN)
Some of values equal 0 → without detection threshold or other informations we use only arithmetic scale (FB and NMSE)
More than a factor 2 for CEA and JAEA (random and systematic errors)
Only about 30% value are within a factor of 2 of observations

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.16</td>
<td>0.31</td>
<td>0.858</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.13</td>
<td>0.26</td>
<td>0.30</td>
<td>0.818</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.42</td>
<td>0.15</td>
<td>0.36</td>
<td>0.912</td>
</tr>
</tbody>
</table>
61 experiments

3 models (CEA, JAEA, IFIN)

Some of values equal 0 → without detection threshold or other informations we use only arithmetic scale (FB and NMSE)

More than a factor 2 for CEA and JAEA (random and systematic errors)

Only about 30% value are within a factor of 2 of observations

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.16</td>
<td>0.31</td>
<td>0.858</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.13</td>
<td>0.26</td>
<td>0.30</td>
<td>0.818</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.42</td>
<td>0.15</td>
<td>0.36</td>
<td>0.912</td>
</tr>
</tbody>
</table>
All models tend to underestimate activity in leaf (less than a factor of 2)
Surely due to very low values

95% confidence limits for FB
TFWT / Tritium model (CERES)
TFWT / Tritium model (JAEA)

+/- a factor-of-two mean bias for prediction

FB (with 95% conf. int.) (Fractionnal Bias)

Overprediction
Underprediction
All models tend to underestimate activity in leaf (less than a factor of 2)
Surely due to very low values

95% confidence limits for FB
TFWT / Tritium model (CERES)
TFWT / Tritium model (JAEA)

+/- a factor-of-two mean bias for prediction

FB (with 95% conf. int.) (Fractionnal Bias)

Overprediction
Underprediction

NMSE (normalized mean square error)
IFIN and JAEA seems make underprediction OBT at the end of harvest but how much?

Difficult to say which model is better

- OBT grain (CERES)
- OBT grain (IFIN)
- OBT grain (JAEA)

Predicted (KBq.kg⁻¹)

Measured (KBq.kg⁻¹)
IFIN and JAEA seems make underprediction OBT at the end of harvest but how much?

Difficult to say which model is better

**Predicted (KBq.kg\(^{-1}\))**

- OBT grain (CERES)
- OBT grain (IFIN)
- OBT grain (JAEA)

**Measured (KBq.kg\(^{-1}\))**

The graph shows a scatter plot comparing predicted and measured values of OBT in grain for different models.
- **14 experiments at the end or harvest**
- 3 models (CEA, JAEA, IFIN)
- Use arithmetic and logarithmic scale → gives about the same results)
- More than a factor 2 for all models (random and systematic errors)
- All model made underprediction (more than a factor of 2 for JAEA)

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.8</td>
<td>1.0</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>VG (1.6)</th>
<th>MG (2.0 or 0.5)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>2.1</td>
<td>1.8</td>
<td>0.5</td>
<td>0.76</td>
</tr>
<tr>
<td>JAEA</td>
<td>15.2</td>
<td>4.0</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>IFIN</td>
<td>1.8</td>
<td>1.9</td>
<td>0.5</td>
<td>0.89</td>
</tr>
</tbody>
</table>
14 experiments at the end or harvest
3 models (CEA, JAEA, IFIN)
Use arithmetic and logarithmic scale → gives about the same results
More than a factor 2 for all models (radom and systematic errors)
All model made underprediction (more than a factor of 2 for JAEA)

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.8</td>
<td>1.0</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>VG (1.6)</th>
<th>MG (2.0 or 0.5)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>2.1</td>
<td>1.8</td>
<td>0.5</td>
<td>0.76</td>
</tr>
<tr>
<td>JAEA</td>
<td>15.2</td>
<td>4.0</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>IFIN</td>
<td>1.8</td>
<td>1.9</td>
<td>0.5</td>
<td>0.89</td>
</tr>
</tbody>
</table>
14 experiments at the end of harvest
3 models (CEA, JAEA, IFIN)
- Use arithmetic and logarithmic scale → gives about the same results
- More than a factor 2 for all models (random and systematic errors)
- All models made underprediction (more than a factor of 2 for JAEA)

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.8</td>
<td>1.0</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>VG (1.6)</th>
<th>MG (2.0 or 0.5)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>2.1</td>
<td>1.8</td>
<td>0.5</td>
<td>0.76</td>
</tr>
<tr>
<td>JAEA</td>
<td>15.2</td>
<td>4.0</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>IFIN</td>
<td>1.8</td>
<td>1.9</td>
<td>0.5</td>
<td>0.89</td>
</tr>
</tbody>
</table>
14 experiments at the end or harvest
3 models (CEA, JAEA, IFIN)
Use arithmetic and logarithmic scale → gives about the same results
More than a factor 2 for all models (random and systematic errors)
All model made underprediction (more than a factor of 2 for JAEA)

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.8</td>
<td>1.0</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.66</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>VG (1.6)</th>
<th>MG (2.0 or 0.5)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>2.1</td>
<td>1.8</td>
<td>0.5</td>
<td>0.76</td>
</tr>
<tr>
<td>JAEA</td>
<td>15.2</td>
<td>4.0</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>IFIN</td>
<td>1.8</td>
<td>1.9</td>
<td>0.5</td>
<td>0.89</td>
</tr>
</tbody>
</table>
14 experiments at the end or harvest
3 models (CEA, JAEA, IFIN)
Use arithmetic and logarithmic scale → gives about the same results
More than a factor 2 for all models (random and systematic errors)
All model made underprediction (more than a factor of 2 for JAEA)

<table>
<thead>
<tr>
<th>Model/Performance (factor 2)</th>
<th>NMSE (0.5)</th>
<th>FB (±2/3)</th>
<th>FAC2</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td>CEA</td>
<td>0.7</td>
<td>0.4</td>
<td>0.5</td>
<td>0.41</td>
</tr>
<tr>
<td>JAEA</td>
<td>1.8</td>
<td>1.0</td>
<td>0.07</td>
<td>0.86</td>
</tr>
<tr>
<td>IFIN</td>
<td>0.8</td>
<td>0.7</td>
<td>0.5</td>
<td>0.66</td>
</tr>
<tr>
<td>Model/Performance (factor 2)</td>
<td>VG (1.6)</td>
<td>MG (2.0 or 0.5)</td>
<td>FAC2</td>
<td>R</td>
</tr>
<tr>
<td>CEA</td>
<td>2.1</td>
<td>1.8</td>
<td>0.5</td>
<td>0.76</td>
</tr>
<tr>
<td>JAEA</td>
<td>15.2</td>
<td>4.0</td>
<td>0.07</td>
<td>0.61</td>
</tr>
<tr>
<td>IFIN</td>
<td>1.8</td>
<td>1.9</td>
<td>0.5</td>
<td>0.89</td>
</tr>
</tbody>
</table>
CEA and IFIN models tend to underestimate activity in leaf (less than a factor of 2), JAEA underestimates about a factor of 3

Surely due to very low values

95% confidence limits for FB
OBT grain / Tritium model (CERES)
OBT grain / Tritium model (IFIN)
OBT grain / Tritium model (JAEA)
+/− a factor-of-two mean bias for prediction

FB (with 95% conf. int.) (Fractionnal Bias)

Overprediction
Underprediction
CEA and IFIN models tend to underestimate activity in leaf (less than a factor of 2), JAEA underestimates about a factor of 3

Surely due to very low values
CEA and IFIN models tend to underestimate activity in leaf (less than a factor of 2), JAEA underestimates about a factor of 4.

Random scatter is less than a factor of 3 (CEA, IFIN) and 5 (JAEA).
CEA and IFIN models tend to underestimate activity in leaf (less than a factor of 2), JAEA underestimates about a factor of 4.

Random scatter is less than a factor of 3 (CEA, IFIN) and 5 (JAEA).
1. Statistical performance measure

2. Simple statistical analysis on wheat experiments

3. Conclusions
CONCLUSIONS (1/2)

- **Statistical analysis can seriously help the models comparison**
  - Performance measures have to be used to compare predictions to observations
  - In case of wheat all models have systematic errors
  - HTO modelling in wheat leaf seems good for the 3 models
  - Systematic errors: \( \frac{C_p}{C_0} = 0.76(\text{JAEA}) \ 0.86(\text{IFIN}& \text{CEA}) \)
  - OBT modelling in wheat grain seems make underprediction for all model
  - Systematic errors: \( \frac{C_p}{C_0} = 0.3(\text{JAEA}) \ 0.48(\text{IFIN}) \ 0.7(\text{CEA}) \)
- Statistical analysis can seriously help the models comparison
- Performance measures have to be used to compare predictions to observations
- In case of wheat all models have systematic errors
- HTO modelling in wheat leaf seems good for the 3 models
- Systematic errors: \( \frac{C_p}{C_o} = 0.76(\text{JAEA}) \ 0.86(\text{IFIN}&\text{CEA}) \)
- OBT modelling in wheat grain seems make underprediction for all models
- Systematic errors: \( \frac{C_p}{C_o} = 0.3(\text{JAEA}) \ 0.48(\text{IFIN}) \ 0.7(\text{CEA}) \)
CONCLUSIONS (1/2)

- Statistical analysis can seriously help the models comparison
- Performance measures have to be used to compare predictions to observations
- In case of wheat all models have systematic errors
  - HTO modelling in wheat leaf seems good for the 3 models
  - Systematic errors: $\left( \frac{C_p}{C_o} = 0.76(JAEA) \ 0.86(IFIN&CEA) \right)$
  - OBT modelling in wheat grain seems make underprediction for all model
  - Systematic errors: $\left( \frac{C_p}{C_o} = 0.3(JAEA) \ 0.48(IFIN) \ 0.7(CEA) \right)$
CONCLUSIONS (1/2)

- Statistical analysis can seriously help the models comparison
- Performance measures have to be used to compare predictions to observations
- In case of wheat all models have systematic errors
- HTO modelling in wheat leaf seems good for the 3 models

Systematic errors:
\[ \frac{C_p}{C_o} = 0.76 \text{ (JAEA)} \ 0.86 \text{ (IFIN & CEA)} \]

- OBT modelling in wheat grain seems make underprediction for all model

Systematic errors:
\[ \frac{C_p}{C_o} = 0.3 \text{ (JAEA)} \ 0.48 \text{ (IFIN)} \ 0.7 \text{ (CEA)} \]
Statistical analysis can seriously help the models comparison
Performance measures have to be used to compare predictions to observations
In case of wheat all models have systematic errors
HTO modelling in wheat leaf seems good for the 3 models
Systematic errors: \[
\frac{C_p}{C_0} = 0.76\text{(JAEA)} 0.86\text{(IFIN & CEA)}
\]
OBT modelling in wheat grain seems make underprediction for all model
Systematic errors: \[
\frac{C_p}{C_0} = 0.3\text{(JAEA)} 0.48\text{(IFIN)} 0.7\text{(CEA)}
\]
CONCLUSIONS (1/2)

- Statistical analysis can seriously help the models comparison
- Performance measures have to be used to compare predictions to observations
- In case of wheat all models have systematic errors
- HTO modelling in wheat leaf seems good for the 3 models
- Systematic errors: \( \frac{C_p}{C_o} = 0.76 \text{(JAEA)} \ 0.86 \text{(IFIN & CEA)} \)
- OBT modelling in wheat grain seems make underprediction for all model
- Systematic errors: \( \frac{C_p}{C_o} = 0.3 \text{(JAEA)} \ 0.48 \text{(IFIN)} \ 0.7 \text{(CEA)} \)
Statistical analysis can seriously help the models comparison
Performance measures have to be used to compare predictions to observations
In case of wheat all models have systematic errors
HTO modelling in wheat leaf seems good for the 3 models
Systematic errors: \( \left( \frac{C_p}{C_o} = 0.76 \text{(JAEA)} \ 0.86 \text{(IFIN&CEA)} \right) \)
OBT modelling in wheat grain seems make underprediction for all model
Systematic errors: \( \left( \frac{C_p}{C_o} = 0.3 \text{(JAEA)} \ 0.48 \text{(IFIN)} \ 0.7 \text{(CEA)} \right) \)
## ARE MODELS IN ACCEPTANCE CRITERIA

### HTO Leaf

<table>
<thead>
<tr>
<th>Test/models</th>
<th>CEA</th>
<th>IFIN</th>
<th>JAEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAC2 &gt; 0.5</td>
<td>no</td>
<td>no</td>
<td>no</td>
</tr>
<tr>
<td>Mean bias within ±30% of the mean ($</td>
<td>FB</td>
<td>&lt; 0.3$ or $0.7 &lt; MG &lt; 1.3$)</td>
<td>ok</td>
</tr>
<tr>
<td>Random scatter ($NMSE &lt; 1.5$ or $VG &lt; 4$)</td>
<td>ok</td>
<td>ok</td>
<td>ok</td>
</tr>
<tr>
<td>Acceptance</td>
<td>ok ?</td>
<td>ok ?</td>
<td>ok ?</td>
</tr>
</tbody>
</table>

### OBT Grain

<table>
<thead>
<tr>
<th>Test/models</th>
<th>CEA</th>
<th>IFIN</th>
<th>JAEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAC2 &gt; 0.5</td>
<td>ok</td>
<td>ok</td>
<td>no</td>
</tr>
<tr>
<td>Mean bias within ±30% of the mean ($</td>
<td>FB</td>
<td>&lt; 0.3$ or $0.7 &lt; MG &lt; 1.3$)</td>
<td>no</td>
</tr>
<tr>
<td>Random scatter ($NMSE &lt; 1.5$ or $VG &lt; 4$)</td>
<td>ok</td>
<td>ok</td>
<td>no</td>
</tr>
<tr>
<td>Acceptance</td>
<td>ok ?</td>
<td>ok ?</td>
<td>no</td>
</tr>
</tbody>
</table>