Uncertainty and sensitivity analysis

David Makowski

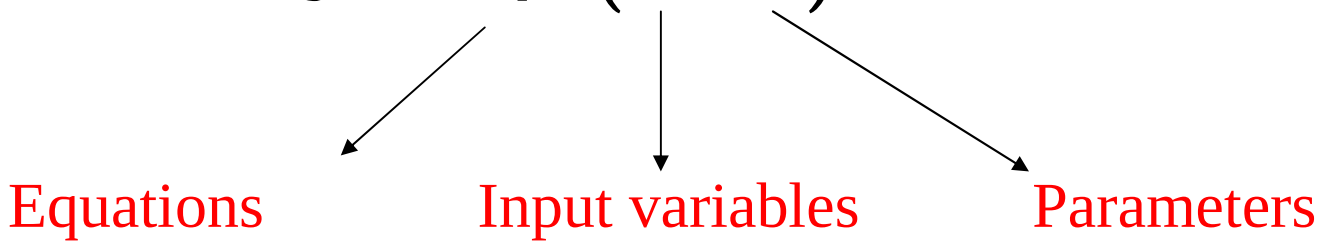
INRA

1. **Introduction**
2. **Uncertainty analysis**
3. **Sensitivity analysis**
4. **Exercises**

1. Introduction

Sources of uncertainty in a model

$$y = f(x, \theta)$$



Notation: z = set of uncertain input variables and parameters

$$z = (z_1, z_2, \ldots, z_p)$$

Types of uncertainty

- ***Lack of knowledge***

Ex: Root crop nitrogen content is not well known

- ***Measurement error***

Ex: Error in disease incidence measurement

- ***Variability of the system characteristics***

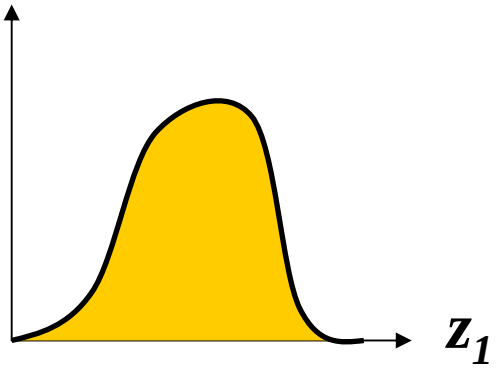
Ex: Variability of « mean daily temperature » between sites and between years

Uncertainty analysis

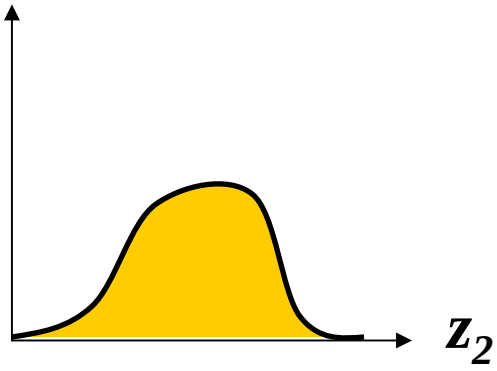
Its purpose is to answer the following question:

« What is the uncertainty about *y*(*z*) resulting from the uncertainty about *z* ? »

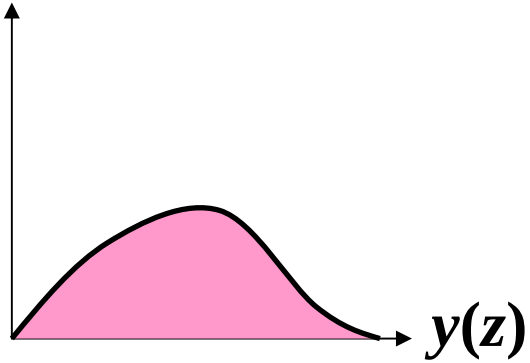
We have



and



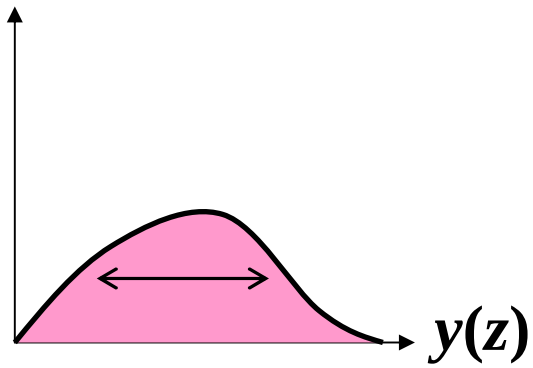
We want to compute



Sensitivity analysis

Its purpose is to answer the question:

« What are the main sources of uncertainty? »



Variance of $y(z)$ = effect of z_1 + effect of z_2 + …

Practical interest

of uncertainty analysis

- Give information about the uncertainty associated with model prediction

- Optimize decision variables

of sensitivity analysis

- Identify the parameters and input variables which strongly influence the model outputs

→ ***Important to know them accurately***

- Identify the parameters and input variables which do not strongly influence the model outputs

→ ***Less important to know them accurately***

Examples of problems where UA and SA are useful

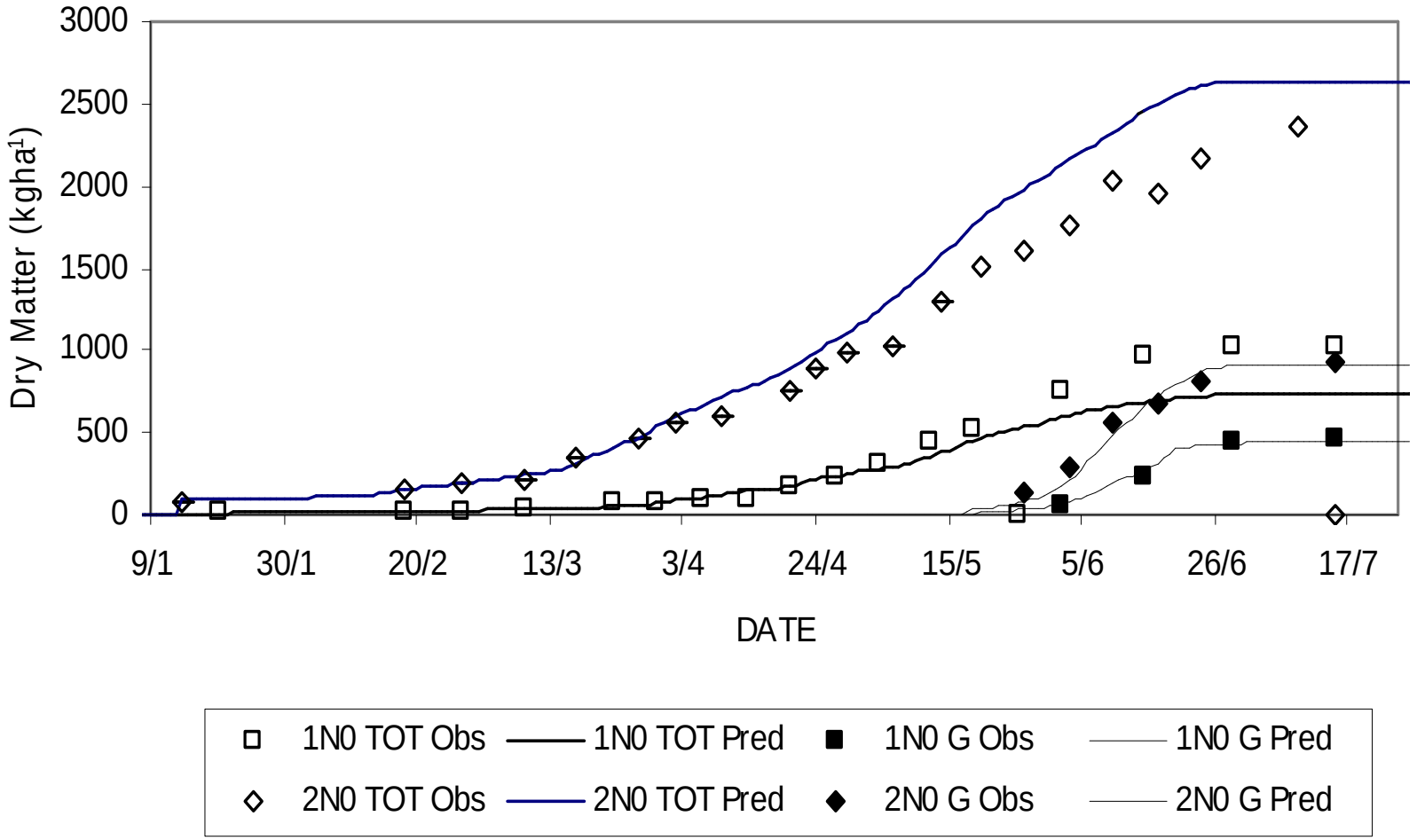
- What is the probability of loosing more than 0.2 t ha^{-1} if the fertilizer dose is reduced by 20%?

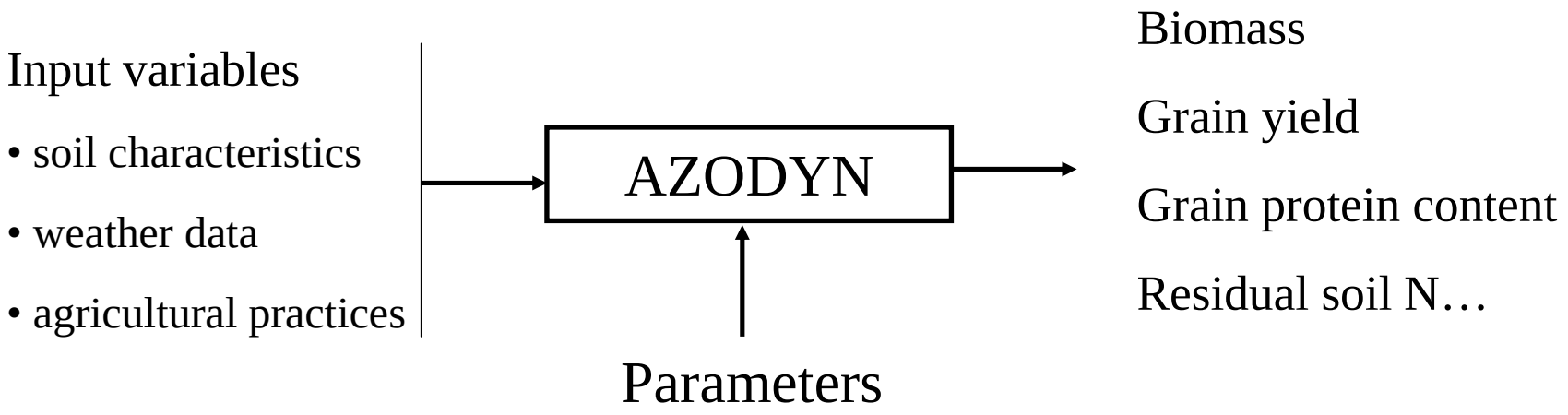
- What is the probability of high disease incidence in a given wheat field when no chemical treatment is applied?

- Is it important to measure soil characteristics for predicting crop yield?

- What are the most important model parameters that need to be estimated genotype by genotype?

Simulations of wheat biomass using the AZODYN dynamic crop model

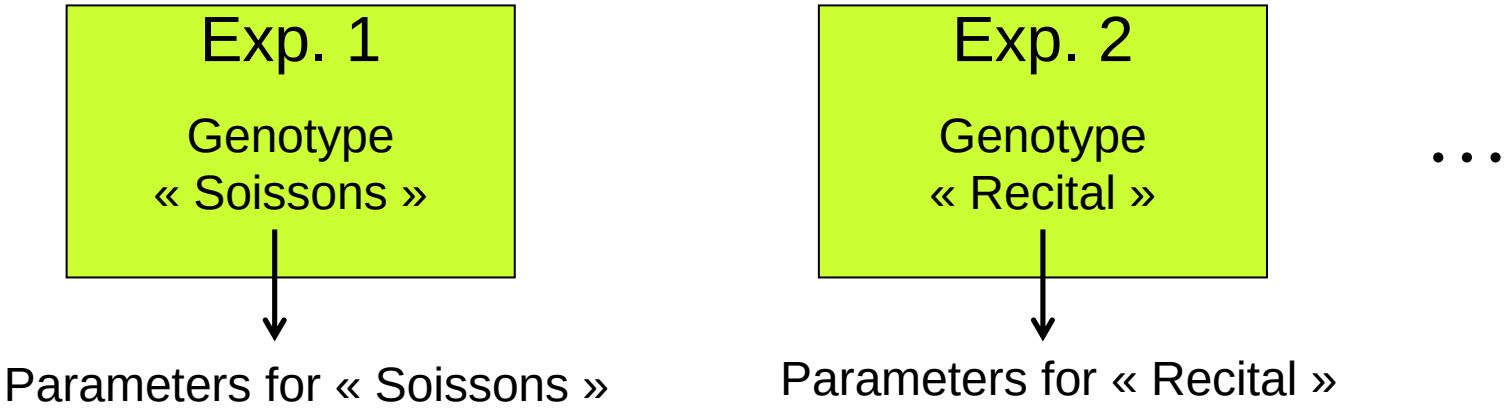



Input variables
• soil characteristics
• weather data
• agricultural practices
AZODYN
Parameters
Biomass
Grain yield
Grain protein content
Residual soil N…

Uncertainty about 13 genotypic parameters

Parameter	Definition	Range	Unit
RDTMAXVAR	Maximal yield	10.0 - 13.7	$t.ha^{-1}$
Ebmax	Radiation use efficiency	2.7-3.3	$g.MJ^{-1}$
D	Ratio of leaf area index to critical nitrogen	0.02-0.045	-
REM2	Fraction of remobilized nitrogen	0.5-0.9	-
K	Extinction coefficient	0.6-0.8	-
Eimax	Ratio of intercepted to incident radiation	0.9-0.99	
Tep.flo	Duration between earing and flowering	100-200	°C.day
R	Ratio of total to above ground nitrogen	1.0-1.5	-
P1GMAXVAR	Maximal weight of one grain	47-65	mg
Lambda	Parameter for calculating nitrogen use efficiency	25-45	-
Mu	Parameter for calculating nitrogen use efficiency	0.6-0.9	-
DJPF	Temperature threshold	150-250	°C.day
NGM2MAXVAR	Maximal grain number	107.95-146.05	-

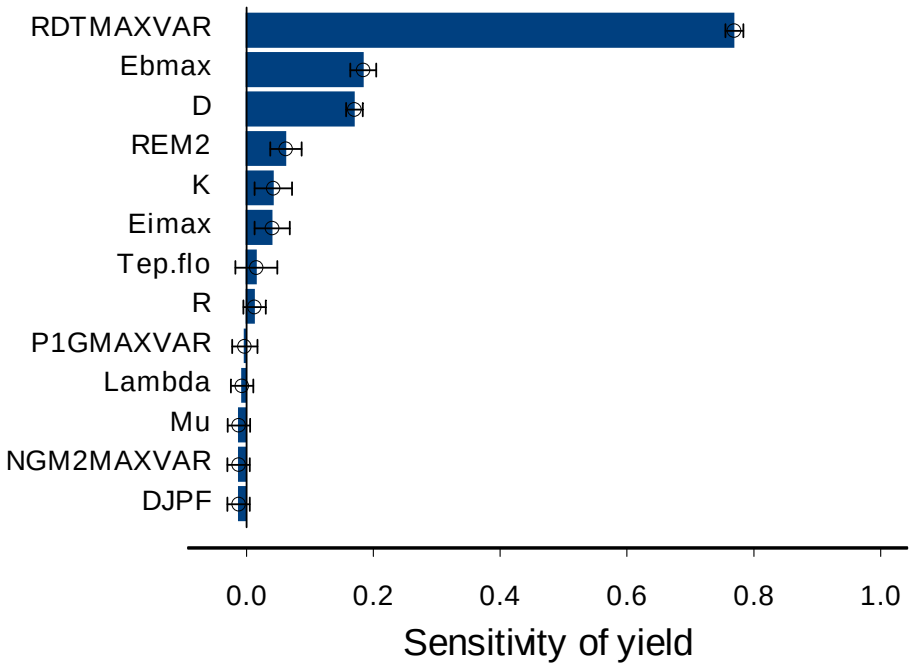
Which parameters should be estimated?



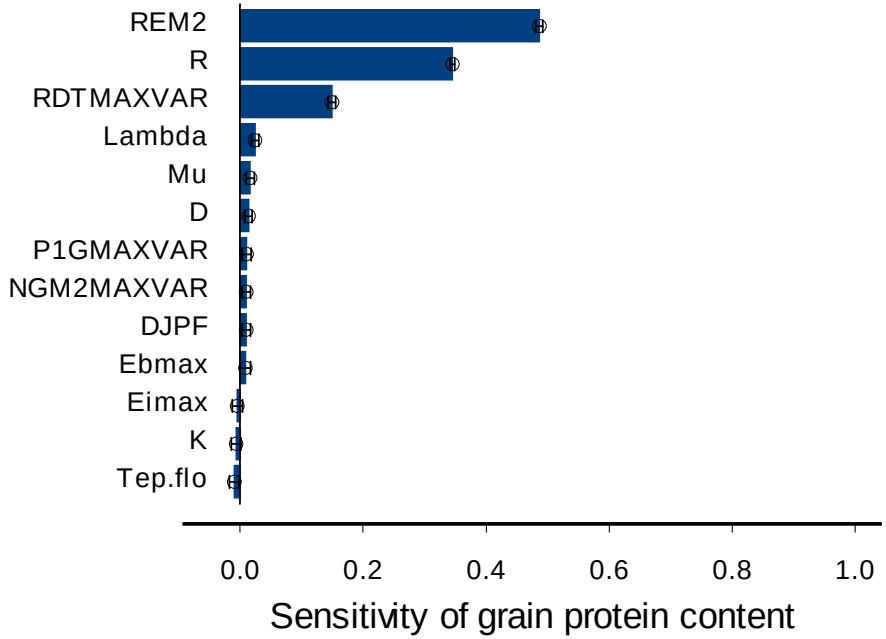
Costly!

Total sensitivity indices for simulated yield and grain protein content

Yield



Grain protein content



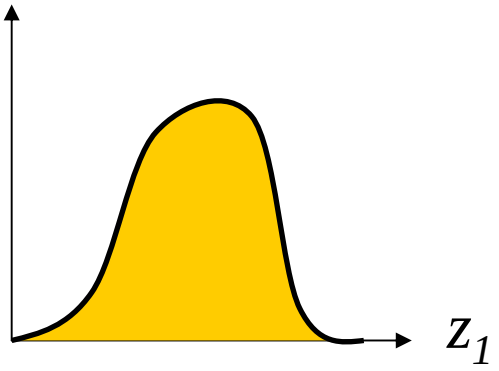
2. Uncertainty analysis

Uncertainty analysis

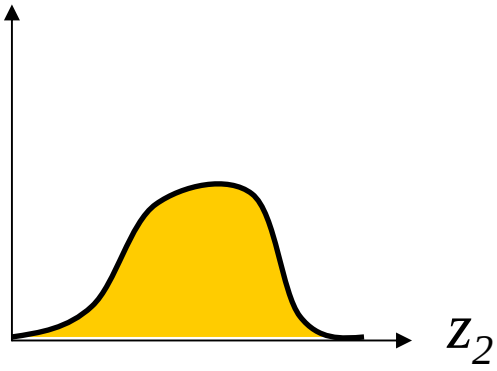
Its purpose is to answer the following question:

« What is the uncertainty about *y*(*z*) resulting from the uncertainty about *z* ? »

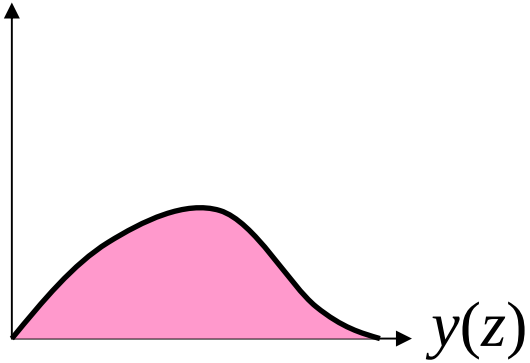
We have



and



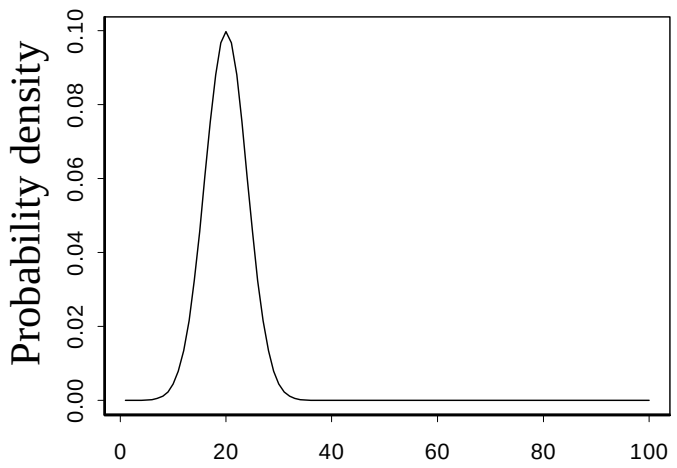
We want to compute



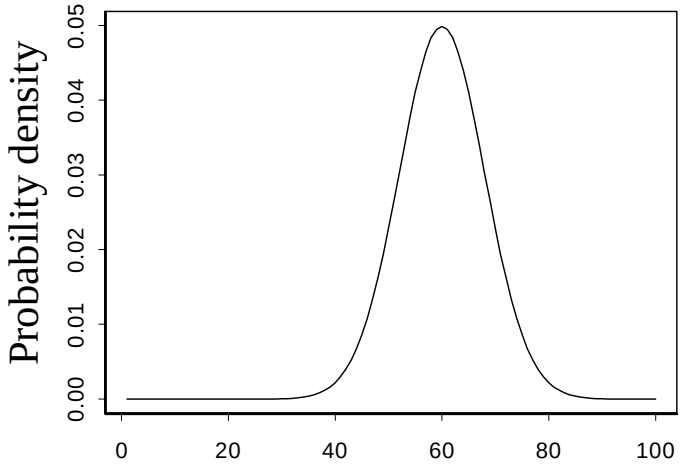
Application to a very simple model

Equation: $y(z_1, z_2) = z_1 + 2\,z_2$

Uncertainty about z_1 and z_2: $z_1 \sim N(20, 16)$ and $z_2 \sim N(60, 64)$



Value of z_1



Value of z_2

Question: Perform an uncertainty analysis

Application to a very simple model

« You need to determine the probability distribution of $y(z_1, z_2)$ from the distributions of z_1 and z_2 » .

Properties:

If z_1 and z_2 are two independant variables with Gaussian distributions then

A z_1 + B z_2 follows a Gaussian distribution

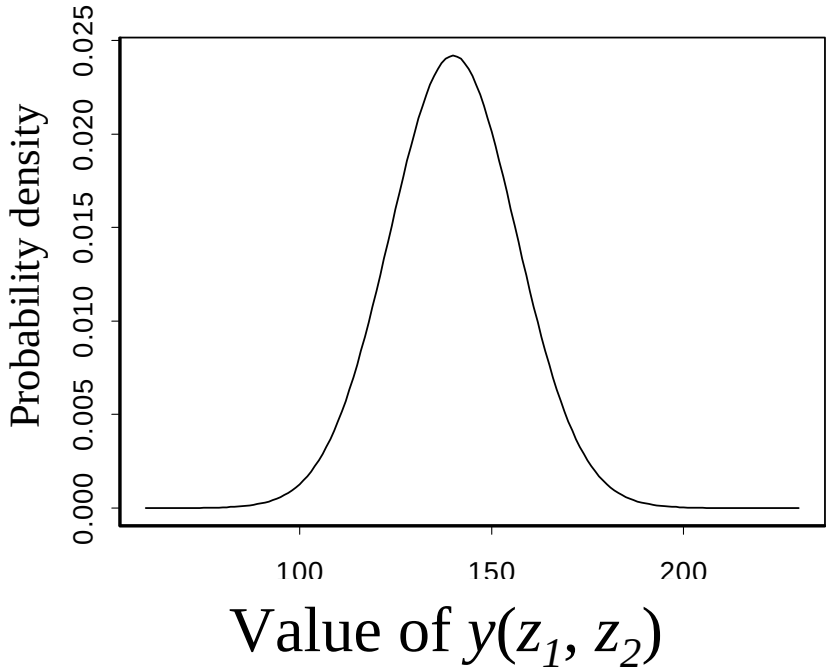
$E(A\ z_1+B\ z_2)=A\ E(z_1)+B\ E(z_2)$

$var(A\ z_1+B\ z_2)=A^2\ var(z_1)+B^2\ var(z_2)$

Application to a very simple model

For this simple model, it is possible to determine the exact expression of the distribution of $y(z_1,z_2)$:

$$y(z_1,z_2) \sim N(140, 272)$$



In general, it is more difficult

• **More complex equations, non linear relationship between *y*(*z*) and *z***

→The analytical expression of the distribution of *y*(*z*) cannot be determined

• **The distribution of *z* is not always well known**

→ Subjective choice

• **Computation times can be long with some models**

→ The number of simulations is limited

Four steps

1. Define the distributions of z_1, …, z_p.

2. Generation of samples from the distributions defined in step 1

3. Computation of $y(z)$ for each generated set z_1, …, z_p

4. Approximate the distribution of $y(z)$

Step 1. Define the distributions

Distributions of the uncertain factors (parameters or input variables) can be defined from

- **Scientific papers or expertise**

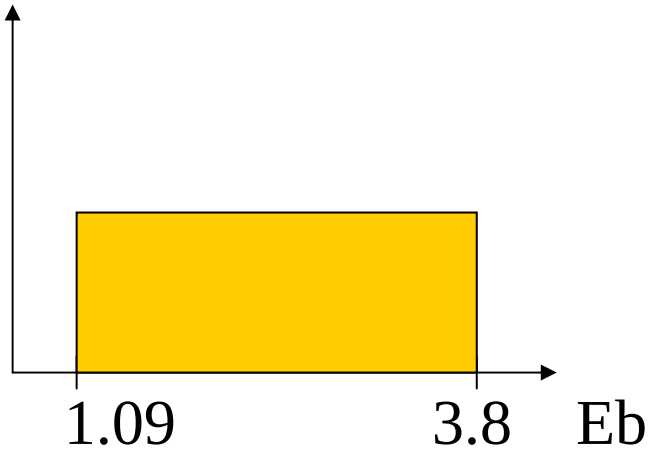
- **Series of data (climatic variables…)**

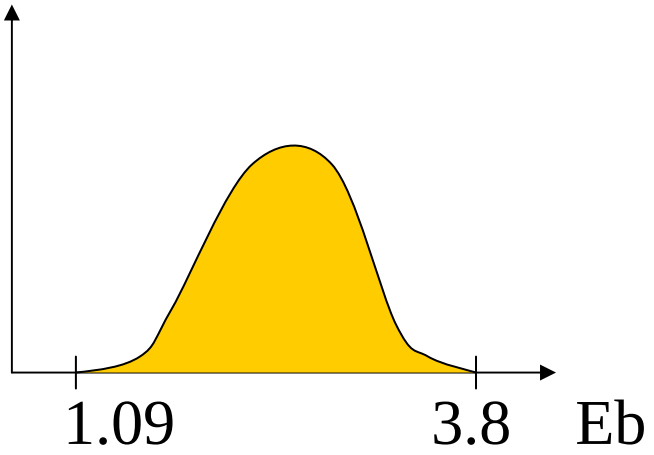
- **Estimated parameter values**

Step 1. Definition the distributions

Example:

from the paper published by Jeuffroy and Recous (1999) in EJA, the intercepted radiation use efficiency varies from **1.09 and 3.8 g MJ^{-1}** for wheat





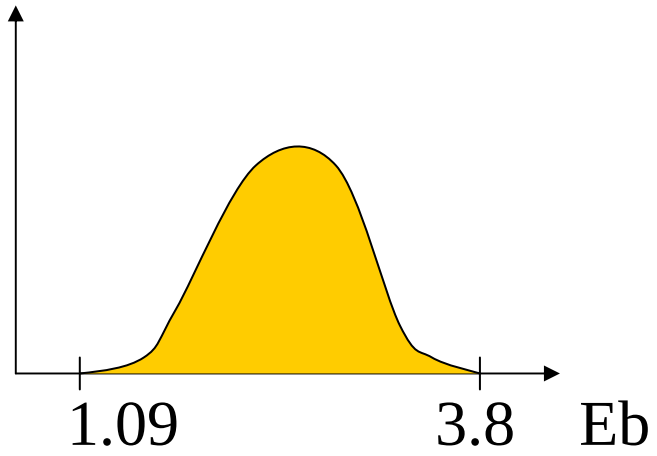
1. Definition of the distributions of z_1, …, z_p.

2. Generation of samples from the distributions defined in step 1

Step 2. Generation of samples from the distributions of z_1, z_2, …

- The number of values of z_1, z_2, …, z_p must be sufficiently high
- Several sampling methods can be used
  - random sampling
  - hypercube latin sampling
  - …
- Many softwares are available to generate values of z_1, z_2, …, z_p

Step 2. Generation of samples from the distributions of z_1, z_2, …



A sample of values of Eb generated from its distribution:

1.2, 1.9, 2.1, 2.2, 2.3, 2.5, 2.7, 3.1, 3.7…

Step 2. Generation of samples from the distributions of z_1, z_2, …

	z_1	z_2	…	z_p
Set 1	1.21	0.85	…	0.99
Set 2	1.97	0.72	…	0.92
…	…	…	…	…
Set N	3.70	0.75	…	0.91

1. Define the probability distributions of z_1, …, z_p

2. Generation of samples from the distributions defined in step 1

3. Computation of $y(z)$ for each generated set z_1, …, z_p

Step 3. Computation of $y(z)$ for each generated set $z_1, \ldots, z_p$

This step can be problematic for complex models due to computational time

Step 3. Computation of $y(z)$ for each generated set $z_1, \ldots, z_p$

	z_1	z_2	…	z_p	y(z)
Set 1	1.21	0.85	…	0.99	90.9
Set 2	1.97	0.72	…	0.92	95.2
…	…	…	…	…	…
Set N	3.70	0.75	…	0.91	81.5

1. Define the probability distributions of z_1, …, z_p.

2. Generation of samples from the distributions defined in step 1

3. Computation of $y(z)$ for each generated set z_1, …, z_p

4. Approximate the distribution of $y(z)$

Step 4. Approximation of the distribution of $y(z)$

- **Describe the N values of $y(z)$ computed at step 3**
- **In general, this is easy**
- **Several possible approaches**
  - compute average and variance
  - compute quantiles (quartiles…),
  - histograms,
  - Cumulative distribution,
  - box plot …

Application to the very simple model

- For this model, the 4-step approach is not necessary because it is possible to determine the exact distribution of $y(z_1, z_2)$

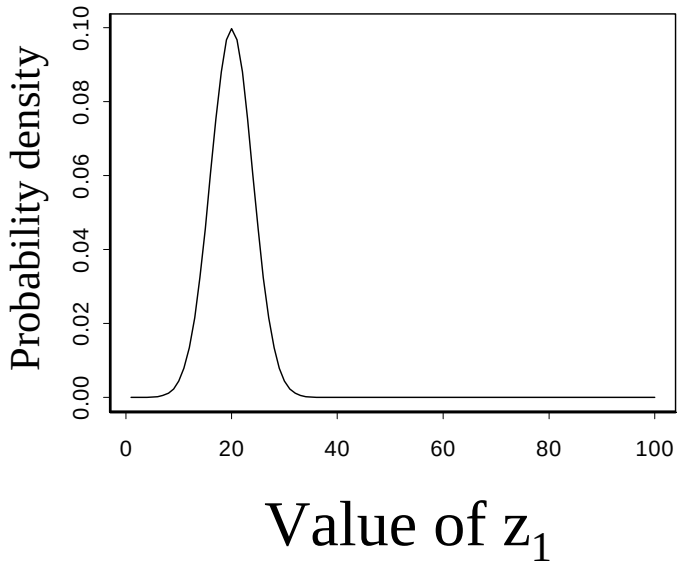
- The 4-step approach is implemented here to show how it works

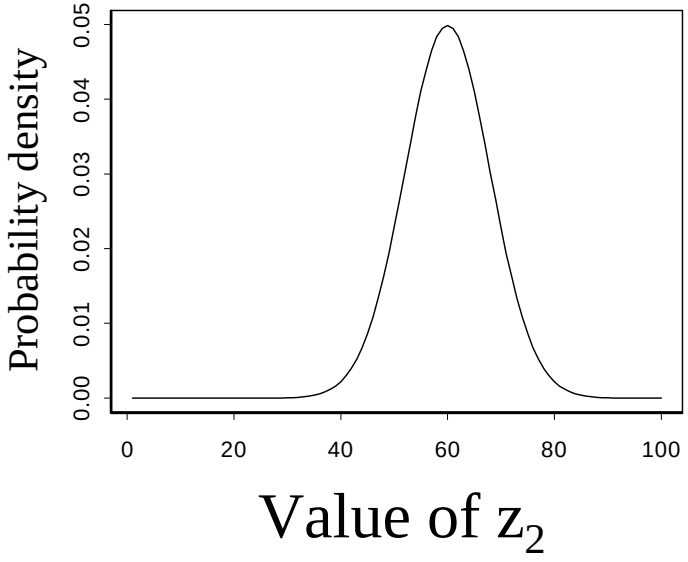
Application to a very simple model

Step 1

Equation : $y(z_1, z_2) = z_1 + 2\, z_2$

Uncertainty about z_1 and z_2 : $z_1 \sim N(20, 16)$, $z_2 \sim N(60, 64)$





Application to a very simple model

Step 2

- N values of z_1 and z_2 are generated

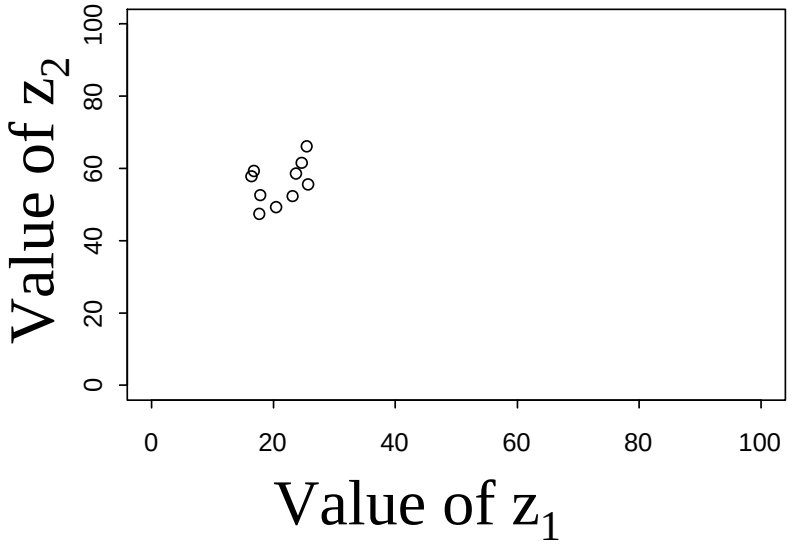
- Several values of N are successively used

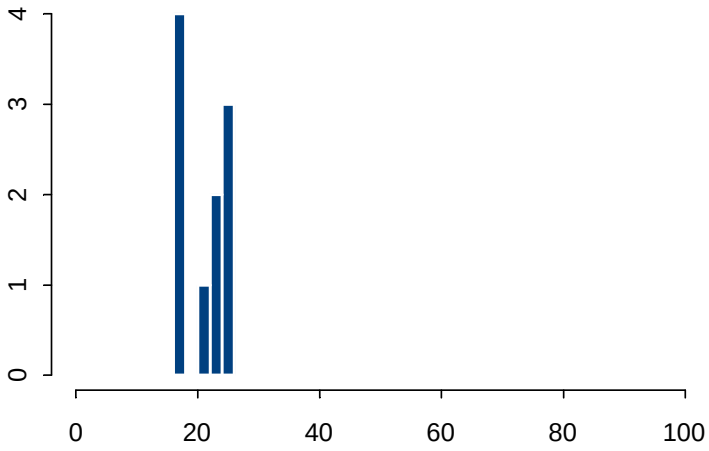
  $N = 10$

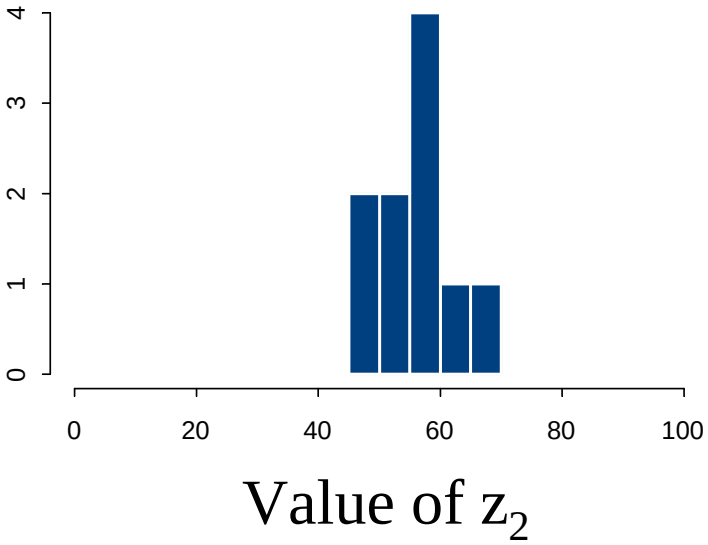
  $N = 100$

  $N = 1000$

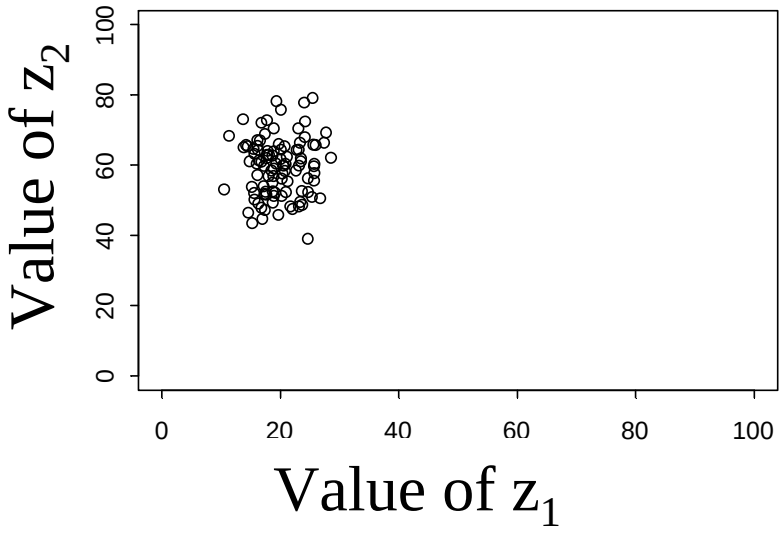
Application. Step 2. N=10

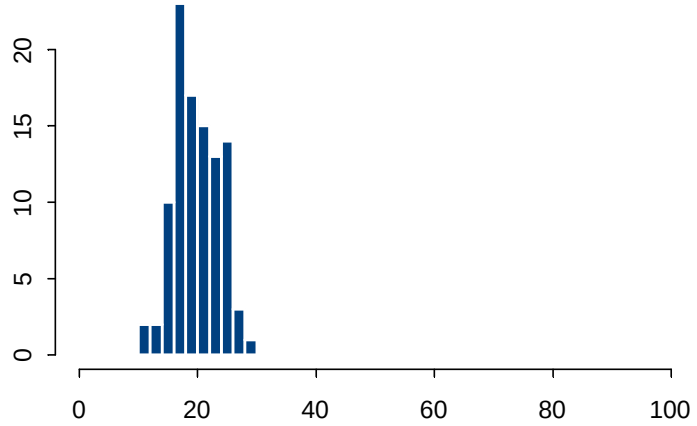


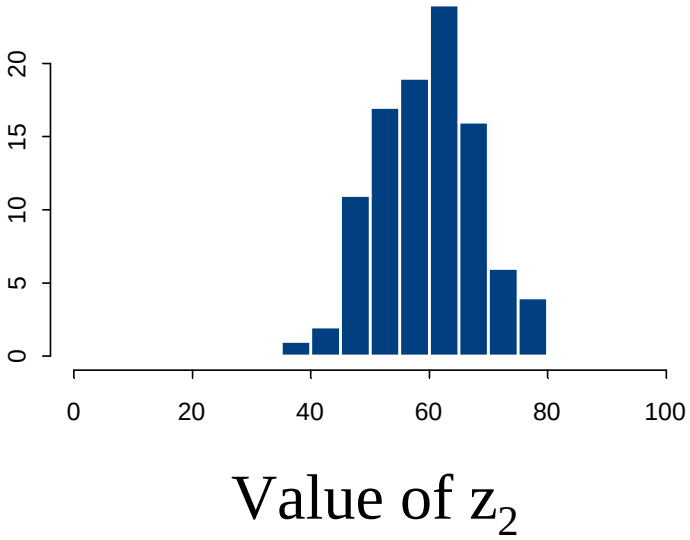




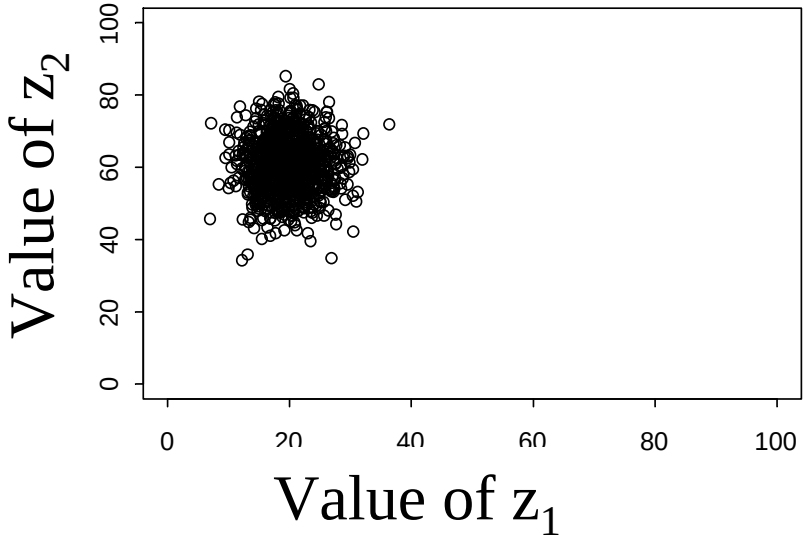
Application. Step 2. N=100

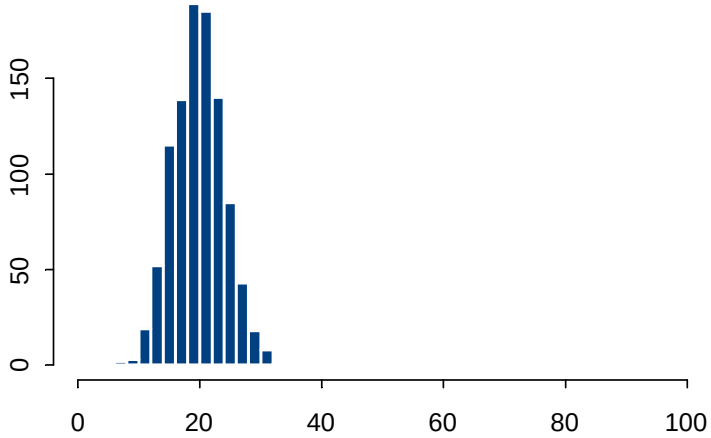


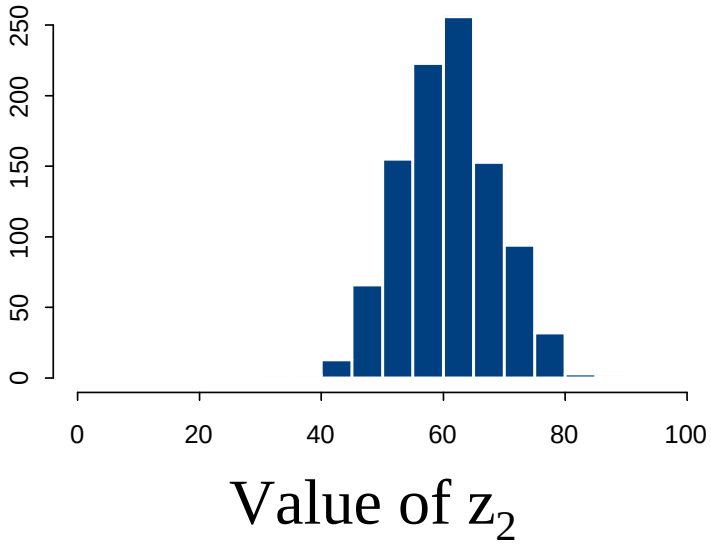




Application. Step 2. N=1000







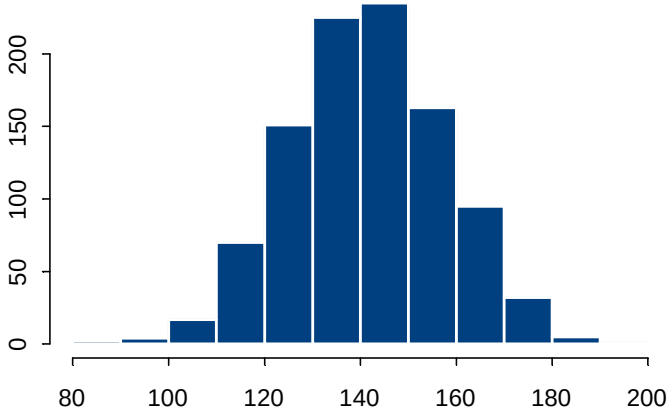
Application. Step 3

z_1	z_2	$y(z_1,z_2)$
16.83	59.30	
23.18	52.33	
16.43	57.85	
20.45	49.25	
25.48	66.11	
25.67	55.53	
24.67	61.55	
17.88	52.58	
23.69	58.54	
17.69	47.38	

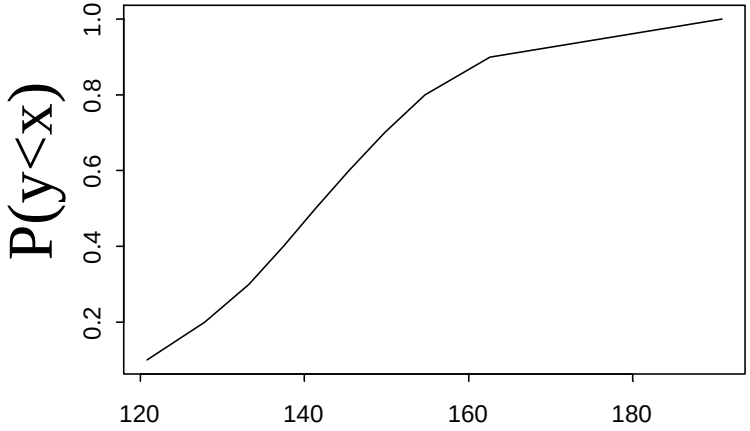
Application. Step 3

z_1	z_2	$y(z_1,z_2)$
16.83	59.30	135.43
23.18	52.33	127.84
16.43	57.85	132.13
20.45	49.25	118.95
25.48	66.11	157.71
25.67	55.53	136.73
24.67	61.55	147.77
17.88	52.58	123.04
23.69	58.54	140.78
17.69	47.38	112.45

Application. Step 4. N=1000



Value of $y(z_1, z_2)$



Value of $y(z_1, z_2)$

Application. Step 4

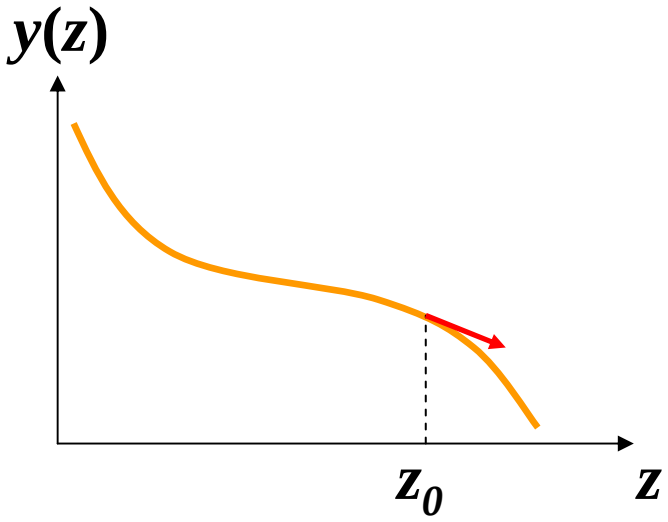
	Mean	Variance	Standard-deviation
N =10	133.28	183.85	13.56
N = 100	138.71	294.96	17.17
N = 1000	141.34	258.23	16.07
N = 5000	139.72	272.51	16.51
N = 7000	139.90	269.45	16.42
True values	140	272	16.49

3. Sensitivity analysis

Local sensitivity analysis or Global sensitivity analysis ?

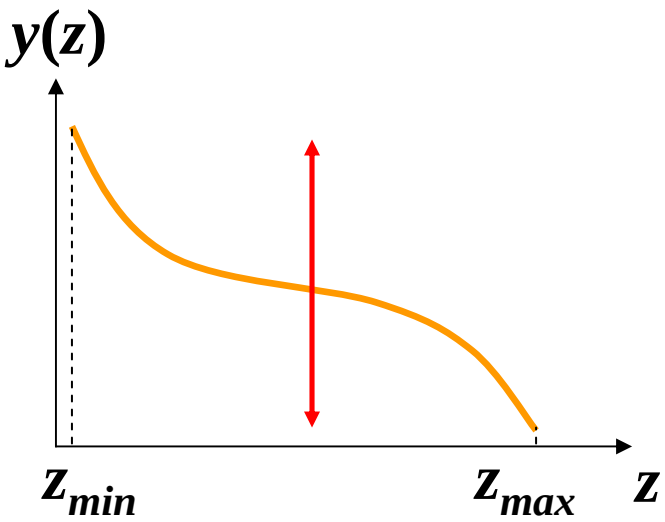
Local SA

variation of $y(z)$ « around » z_0



Global SA

Global variation of $y(z)$ when z varies



Practical interest of sensitivity analysis

i) Identify the parameters and input variables which strongly influence the model outputs

→ ***Important to know them accurately***

ii) Identify the parameters and input variables which do not strongly influence the model outputs

→***Less important to know them accurately***

iii) Analyze the behaviour of the model at some points

Local sensitivity analysis

Based on the computation of derivatives

Global sensitivity analysis

It consists in

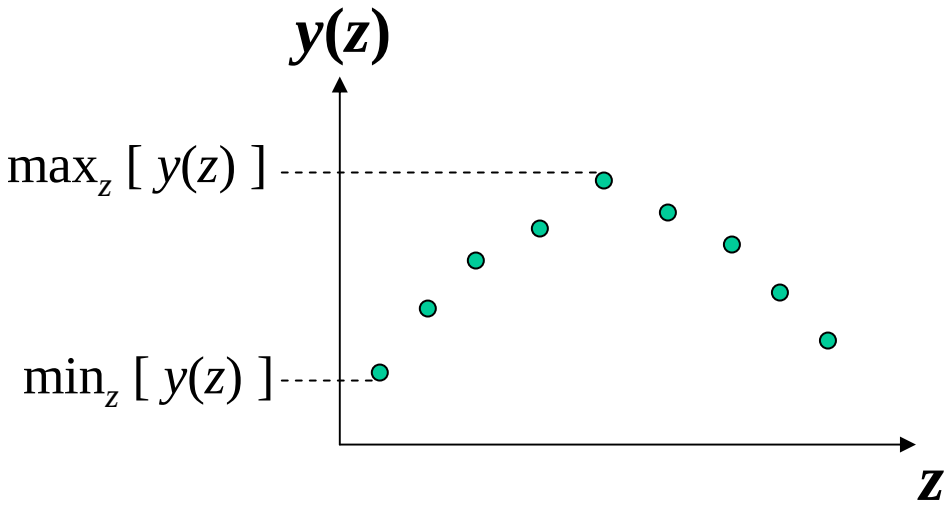
- Defining sensitivity indices
- Compute the indices by varying the uncertain factors $z_1, \ldots, z_p$ over their ranges

A simple sensitivity indice

Bauer and Hamby (1991)

- Define a **series of values** for each uncertain factor
- Fixe **all factors but *zi*** to reference values
- Compute the index for **the factor *zi*** as follows

$$\mathbf{I_{zi} = \{ \max_{zi} [\, y(z) \,] - \min_{zi} [\, y(z) \,] \} / \max_{zi} [\, y(z) \,]}$$



Application

Equation: $y(z_1, z_2) = z_1 + 2\ z_2$

Define five values for z_2: 40, 50, 60, 70, 80.

Fixe z_1 to 20.

Question: What is the value of the Bauer-Hamby index for z_2?

Application

$\max_{z2}$ [$y(z_1=20, z_2)$] = 20 + 2*80 = 180

$\min_{z2}$ [$y(z_1=20, z_2)$] = 20 + 2*40 = 100

I_{z2} = (180 - 100) / 180 = 0.444

Limitation of the Bauer-Hamby index

- Each factor is analyzed separately
- The index value may depend on the reference values

Example:

$y(z_1, z_2, z_3) = z_1 + 2*z_2*z_3$.

$I_{z2} = 0$ si $z_3=0$.

$I_{z2} \neq 0$ si $z_3 \neq 0$.

Interactions between factors are not taken into account

Variance-based sensitivity indices

$$\mathrm{Var}[\, y(z)\,] = V_{z1} + V_{z2} + V_{z3} + \ldots + V_{z1.z2} + V_{z1.z3} + \ldots$$

Total variance of the output variable

Main effects of the uncertain factors

Interactions

First order sensitivity index for $z_1 = V_{z1} / \mathrm{Var}[\, y(z)\,]$

Total sensitivity index for $z_1 = (\, V_{z1} + V_{z1.z2} + V_{z1.z3} + \ldots) / \mathrm{Var}[\, y(z)\,]$

Meaning of the total sensitivity index

- **Total sensitivity index of z_i (IT_i) = Residual fraction of the variance of y when only z_i remains unknown.**

- **IT_i ranges from 0 to 1.**

IT_i close to *0* ⟶ Not necessary to estimate accurately z_i

IT_i close to *1* ⟶ *zi* needs to be accurately estimated.

Global SA = the first three steps of UA + a new step

1. Define the probability distributions of z_1, …, z_p.

2. Generation of samples from the distributions defined in step 1

3. Computation of $y(z)$ for each generated set z_1, …, z_p

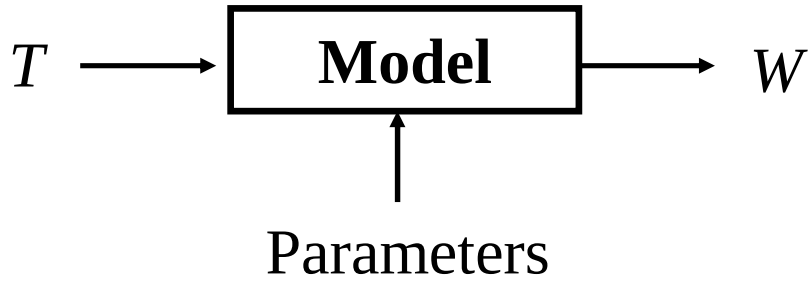
4. Compute sensitivity indices

Ex: Bauer & Hamby, ANOVA, Monte Carlo, Sobol, FAST etc.

Exercises

Generic model for computing the wetness duration requirement for crop infection by fungus

(Magarey *et al*., 2005)



W = wetness duration requirement (h)
T = average temperature (°C)

Generic model for computing the wetness duration requirement for crop infection by fungus

(Magarey *et al*., 2005)

$$W = W_{\min} / f(T)\ ,\ \text{but not higher than}\ W_{\max}$$

$$f(T) = \left(\frac{T_{\max} - T}{T_{\max} - T_{opt}} \right) \left(\frac{T - T_{\min}}{T_{opt} - T_{\min}} \right)^{\left(T_{opt} - T_{\min}\right) / \left(T_{\max} - T_{opt}\right)}$$

Five parameters: T_{min}, T_{opt}, T_{max}, W_{min}, W_{max}

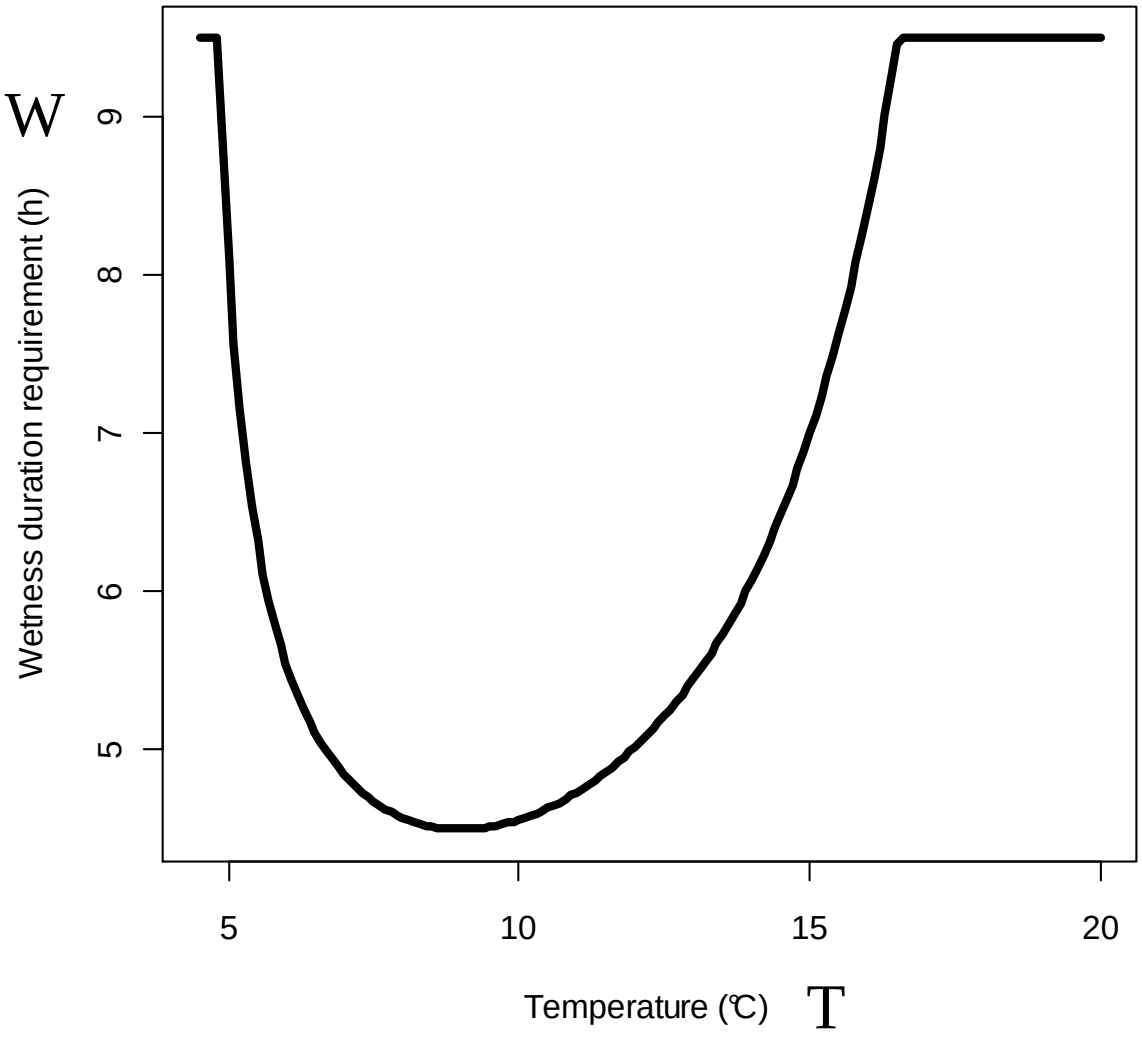
- **Parameters can be estimated from data and from published papers for different fungus**

- **Uncertainty remain about parameter values**

- **Important to analyze the uncertainty about *W* and to identify the most important parameters**

Example of estimated parameter values and corresponding model predictions

Tmin= 4.5 °C, Topt= 9 °C, Tmax=20 °C,
Wmin=4.5 h, Wmax= 9.5 h



Uncertainty about parameter values

		Min	Max
Tmin	(°C):	2.5	6
Topt	(°C):	8	10
Tmax	(°C):	18	22
Wmin	(h):	4	5
Wmax	(h):	8	11

Questions

1. **Perform an uncertainty analysis for *W* using uniform distributions for the uncertain parameters.**

2. Perform a sensitivity analysis for *W* using an ANOVA

1. Perform an uncertainty analysis for *W* using uniform distributions for the uncertain parameters

i. Define the parameter distributions

ii. Generate *N* sets of parameter values (*N*=10, 100, 1000, 2000)

iii. Compute *W* for each set

iv. Describe the distribution of *W*

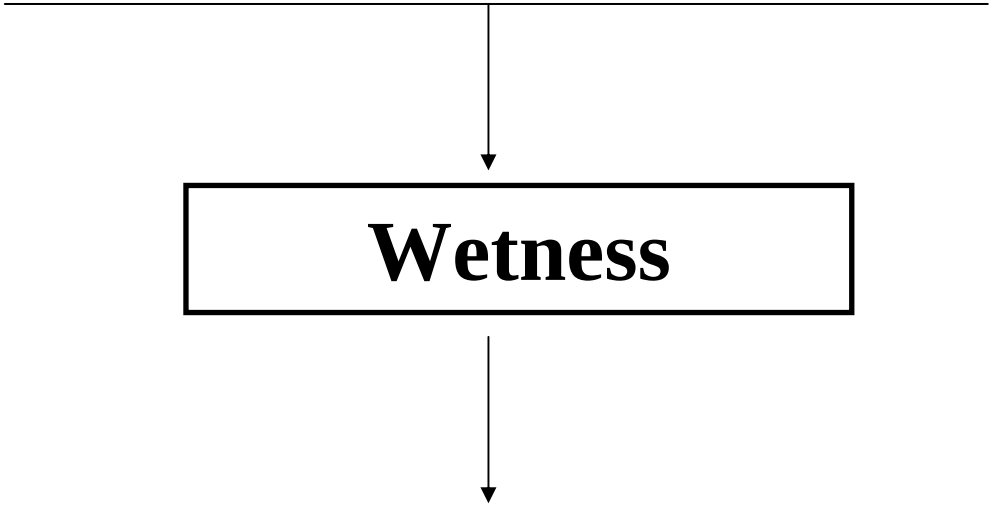
R

```
X <- c(0, 1, 2)                    # the arrow <- means « put into »

                                   # c() can be used to create a vector

Y = c(0, 1, 2)                     # = can be used instead of <-

TAB<-data.frame(X, Y)              #create a table including two columns

print(X)                           # the brackets () are used to define the
                                   # inputs of a R function

plot(X, Y)                         # scatter plot

X[2]                               # the square brackets [] are used to
                                   # define subsets of vector or matrix

print(X[2])                        # return « 1 » here

print(X[Y>0])                      # return « 1   2 »

M<-matrix(nrow=2, ncol=3)          # create a 2 by 3 matrix

M[ , 2]                            # second column of the matrix
```

A R function computing Wetness duration requirement

```
Wetness<-function(T, Tmin, Topt, Tmax, Wmin, Wmax) {

fT<-((Tmax-T)/(Tmax-Topt))*(((T-Tmin)/(Topt-Tmin))
                                ^((Topt-Tmin)/(Tmax-Topt)))

W <- Wmin/fT
W[W>Wmax]<-Wmax
return(W)

}
```

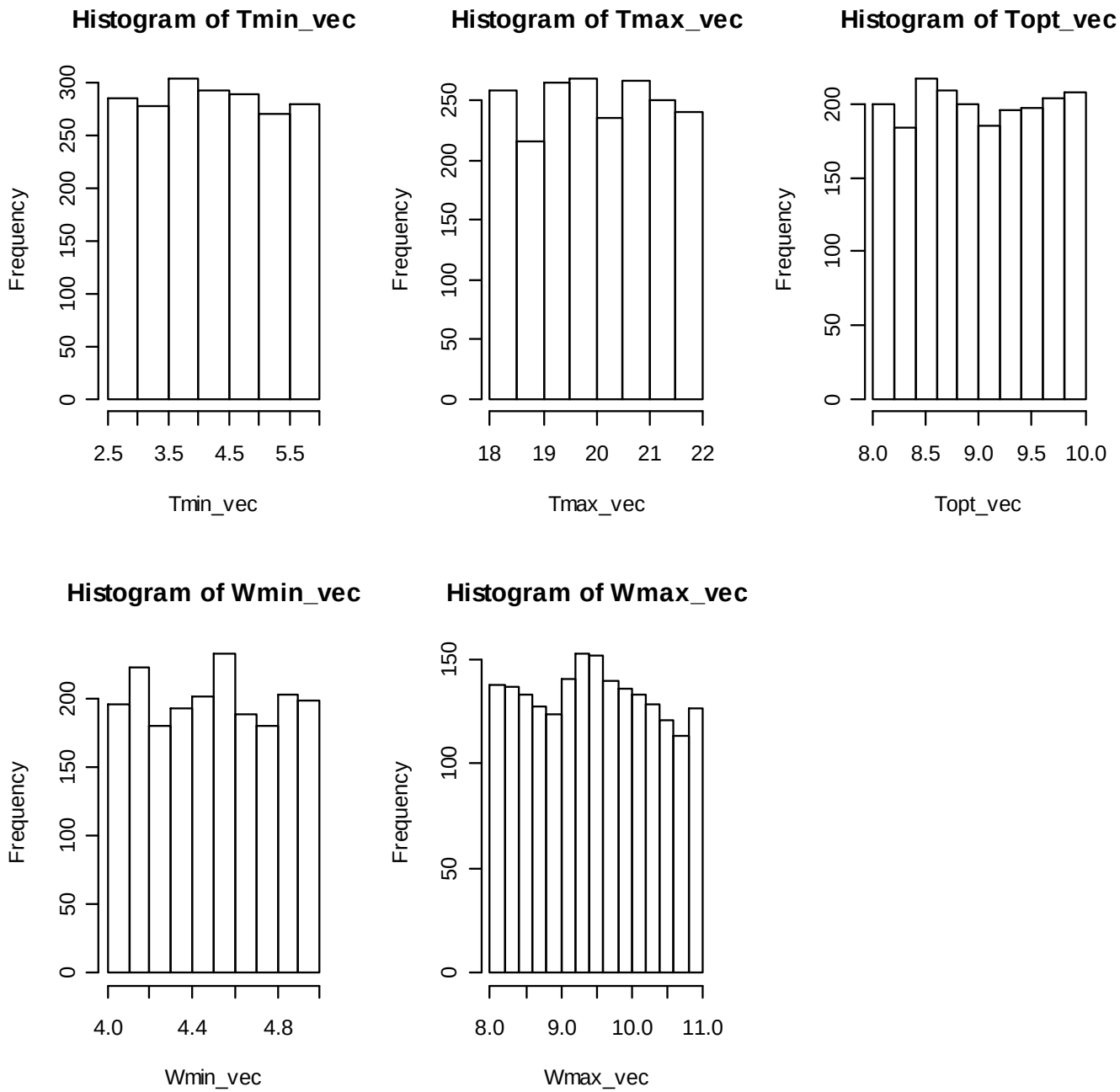

T, Tmin, Topt, Tmax, Wmin, Wmax



Wetness duration requirement

Generation of parameter values

```
Num<-100

Tmin_vec<-runif(Num, 2.5, 6)

Topt_vec<-runif(Num, 8, 10)

Tmax_vec<-runif(Num, 18, 22)

Wmin_vec<-runif(Num, 4, 5)

Wmax_vec<-runif(Num, 8, 11)

plot(Wmin_vec, Wmax_vec)

plot(Tmin_vec, Tmax_vec)
```

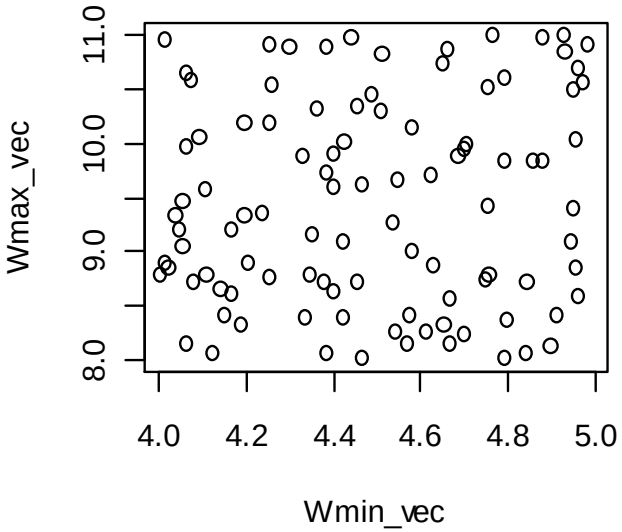


Histogram of Tmin_vec
Frequency
0
50
100
150
200
250
300
2.5
3.5
4.5
5.5
Tmin_vec
Histogram of Tmax_vec
Frequency
0
50
100
150
200
250
18
19
20
21
22
Tmax_vec
Histogram of Topt_vec
Frequency
0
50
100
150
200
8.0
8.5
9.0
9.5
10.0
Topt_vec
Histogram of Wmin_vec
Frequency
0
50
100
150
200
4.0
4.4
4.8
Wmin_vec
Histogram of Wmax_vec
Frequency
0
50
100
150
8.0
9.0
10.0
11.0
Wmax_vec

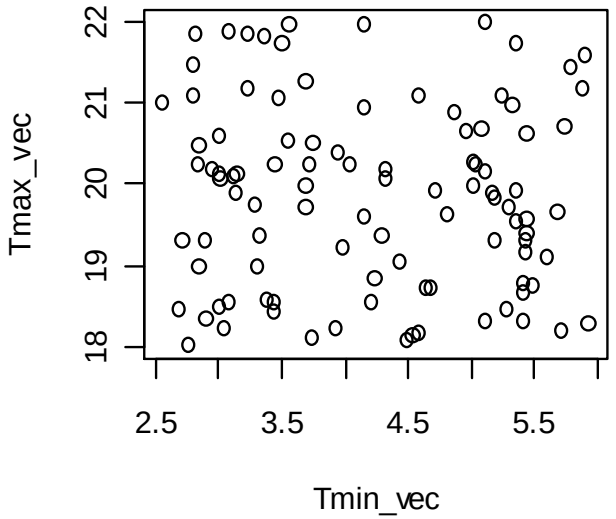
Computation of model outputs

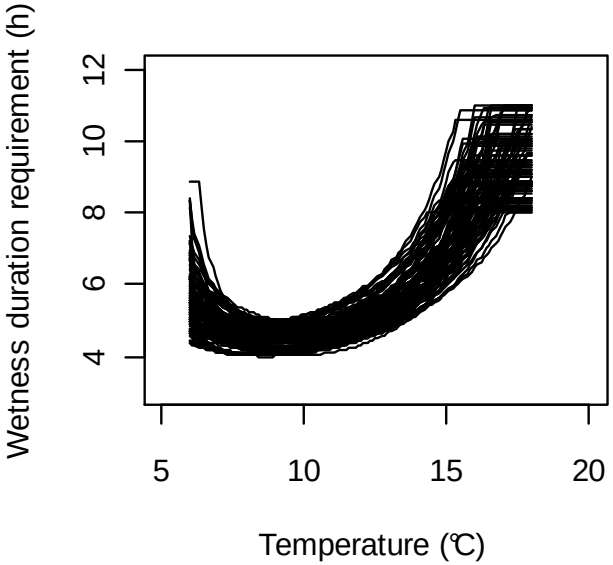
```
T_vec<-seq(from=6, to=18, by=0.1)

W_mat<-matrix(nrow=Num, ncol=length(T_vec))

for (i in 1:Num) {

W_mat[i,]<-Wetness(T_vec, Tmin_vec[i], Topt_vec[i],
                   Tmax_vec[i], Wmin_vec[i], Wmax_vec[i])

lines(T_vec, W_mat[i,])

}
```

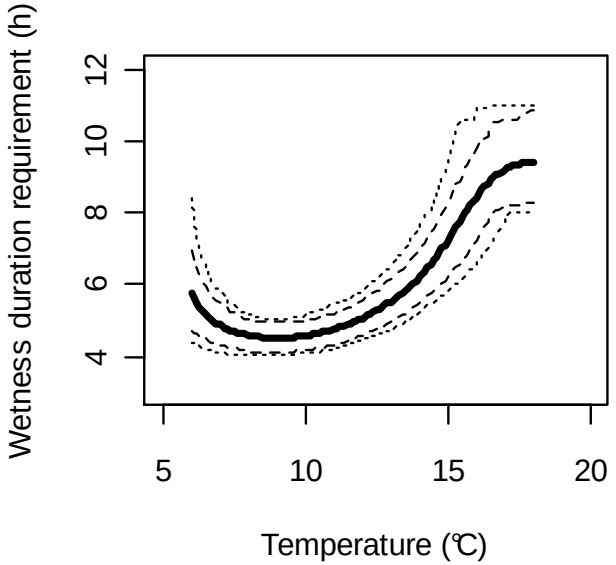
Approximation of the output distribution

```
mean_vec<-apply(W_mat, 2, mean)
Q0.01_vec<-apply(W_mat, 2, quantile, 0.01)
Q0.1_vec<-apply(W_mat, 2, quantile, 0.1)
Q0.9_vec<-apply(W_mat, 2, quantile, 0.9)
Q0.99_vec<-apply(W_mat, 2, quantile, 0.99)

plot(c(0), c(0), pch=" ", xlab="Temperature (°C)",
ylab="Wetness duration requirement (h)", xlim=c(5, 20),
ylim=c(3, 12))
lines(T_vec, mean_vec, lwd=3)
lines(T_vec, Q0.9_vec, lty=2)
lines(T_vec, Q0.1_vec, lty=2)
lines(T_vec, Q0.99_vec, lty=9)
lines(T_vec, Q0.01_vec, lty=9)
```


Wmax_vec
8.0
9.0
10.0
11.0
4.0
4.2
4.4
4.6
4.8
5.0
Wmin_vec


Tmax_vec
18
19
20
21
22
2.5
3.5
4.5
5.5
Tmin_vec


Wetness duration requirement (h)
4
6
8
10
12
5
10
15
20
Temperature (℃)


Wetness duration requirement (h)
4
6
8
10
12
5
10
15
20
Temperature (℃)

2. Perform a sensitivity analysis for *W* using ANOVA

i. Define three values for each parameters

ii. Generate all possible combinations of parameter values

iii. Compute *W* for each combination

iv. Perform an ANOVA and compute sensitivity indices

A R function computing Wetness duration requirement

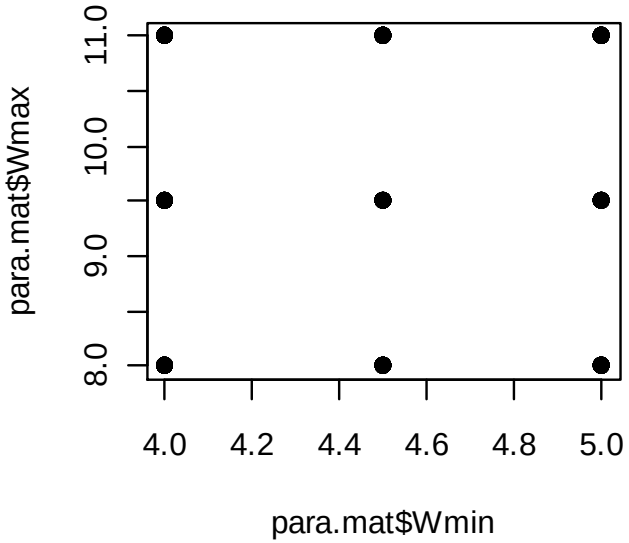
```
Wetness<-function(T, Tmin, Topt, Tmax, Wmin, Wmax) {

fT<-((Tmax-T)/(Tmax-Topt))*(((T-Tmin)/(Topt-Tmin))
                                    ^((Topt-Tmin)/(Tmax-Topt)))

W <- Wmin/fT

W[W>Wmax]<-Wmax

return(W)

}

# The code of the function must be written within { }
```

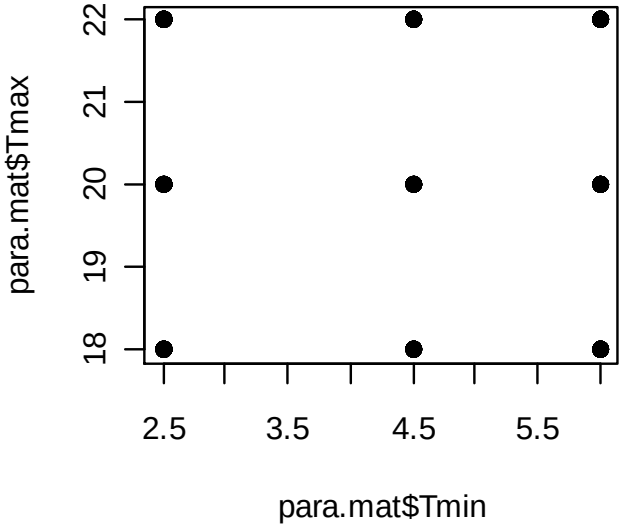
Define several values for each parameters and Generate all possible combinations of parameter values

```
# Create a table including 243 parameter combinations

para.mat<-expand.grid( Tmin=c(2.5, 4.5, 6),
                       Topt=c(8, 9, 10),
                       Tmax=c(18, 20, 22),
                       Wmin=c(4, 4.5, 5),
                       Wmax=c(8, 9.5, 11))

print(para.mat)

plot(para.mat$Wmin, para.mat$Wmax, pch=19)

plot(para.mat$Tmin, para.mat$Tmax, pch=19)
```

Factorial experimental design

	Tmin	Topt	Tmax	Wmin	Wmax
1	2.5	8	18	4.0	8.0
2	4.5	8	18	4.0	8.0
3	6.0	8	18	4.0	8.0
4	2.5	9	18	4.0	8.0
5	4.5	9	18	4.0	8.0
6	6.0	9	18	4.0	8.0
7	2.5	10	18	4.0	8.0
8	4.5	10	18	4.0	8.0
9	6.0	10	18	4.0	8.0
10	2.5	8	20	4.0	8.0
11	4.5	8	20	4.0	8.0
…..					
243					


para.mat$Wmax
11.0
10.0
9.0
8.0
4.0
4.2
4.4
4.6
4.8
5.0
para.mat$Wmin


para.mat$Tmax
22
21
20
19
18
2.5
3.5
4.5
5.5
para.mat$Tmin

Compute *W* for each combination

```
# Temperature values

T.vec <-c(7, 10, 13)

# Create an empty matrix to store the simulated values

W.mat<-matrix(nrow=243, ncol=3)

# Loop for simulating W

for (i in 1:243) {

W.mat[i,]<-Wetness(T.vec, para.mat$Tmin[i], para.mat$Topt[i],
para.mat$Tmax[i], para.mat$Wmin[i], para.mat$Wmax[i])

}
```

Results of the virtual experiments

	W	Tmin	Topt	Tmax	Wmin	Wmax
1	4.215856	2.5	8	18	4	8
2	4.268417	4.5	8	18	4	8
3	4.352753	6	8	18	4	8
4	4.058148	2.5	9	18	4	8
5	4.070403	4.5	9	18	4	8
6	4.088521	6	9	18	4	8
7	4.000000	2.5	10	18	4	8
8	4.000000	4.5	10	18	4	8
9	4.000000	6	10	18	4	8
10	4.163939	2.5	8	20	4	8
11	4.207156	4.5	8	20	4	8
….						
243						

Sensitivity indices

```
#Define the sets of parameter values as factors

Tmin<-as.factor(para.mat$Tmin)
Topt<-as.factor(para.mat$Topt)
Tmax<-as.factor(para.mat$Tmax)
Wmin<-as.factor(para.mat$Wmin)
Wmax<-as.factor(para.mat$Wmax)

#Select the simulations obtained for T=10
W<-W.mat[,2]

#Create a table
TAB<-data.frame(W, Tmin, Topt, Tmax, Wmin, Wmax)
```

Sensitivity indices

```
#ANOVA (sum of squared associated with main effects and interactions)

Fit<-summary(aov(W~Tmin*Topt*Tmax*Wmin*Wmax, data=TAB))
print(Fit)

#Computation of sensitivity indices
SumSq<-Fit[[1]][,2]
Total<-242*var(W)
Indices<-100*SumSq/Total
print(Indices)
TabIndices<-cbind(Fit[[1]],Indices)
print(TabIndices)
TabIndices<-TabIndices[order(Indices, decreasing=T),]
print(TabIndices)
```

	Sum Sq	Mean Sq	Indices
Wmin	4.241487e+01	2.120744e+01	9.291296e+01
Topt	2.762289e+00	1.381145e+00	6.051002e+00
Tmin:Topt	1.267748e-01	3.169371e-02	2.777098e-01
Tmin	1.138496e-01	5.692479e-02	2.493960e-01
Tmax	1.010550e-01	5.052752e-02	2.213686e-01
Topt:Tmax	9.954449e-02	2.488612e-02	2.180597e-01
Topt:Wmin	2.273489e-02	5.683723e-03	4.980249e-02

Some references

Lacroix, A., N. Beaudoin, D. Makowski. 2005. Agricultural water nonpoint pollution control under uncertainty and climate variability. *Ecological Economics* 53:115-127

Makowski, D., C. Naud, M-H. Jeuffroy, A. Barbottin, H. Monod. 2006. Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model predictions. *Reliability Engineering and System Safety* 91:1142-1147.

Magarey RD, Sutton TB, Thayer CL. 2005. A simple generic infection model for foliar fungal plant pathogens. *Phytopathology* 95, 92-100.

Monod, H., C. Naud, D. Makowski. 2006. Uncertainty and sensitivity analysis for crop models. *In: Working with dynamic crop models.* D. Wallach, D. Makowski, J. Jones Eds, Elsevier. p. 55-100.

Saltelli, A., S. Tarantola, F. Campolongo, M. Ratto. 2004. « *Sensitivity analysis in practice, a guide to assessing scientific models* ». Wiley.