Probabilistic Risk Analysis (PRA) and Bayesian Decision Theory (BDT)

USGS CC PDM 1.0

Marcel van Oijen
(Independent scientist,
Edinburgh, U.K.)

Mark Brewer
(Biomathematics and
Statistics Scotland)

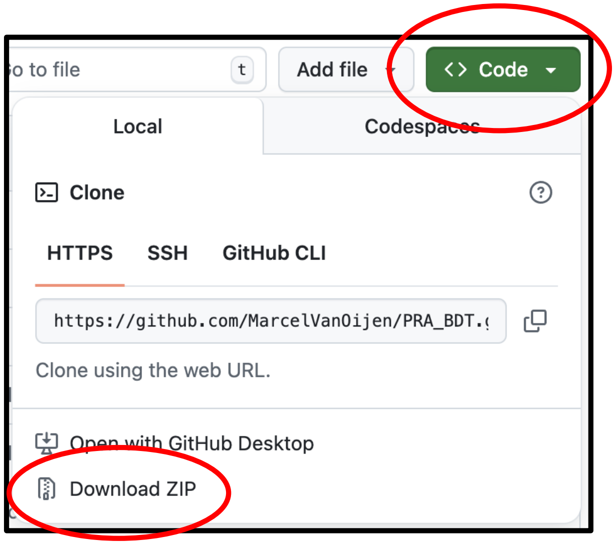
Contents

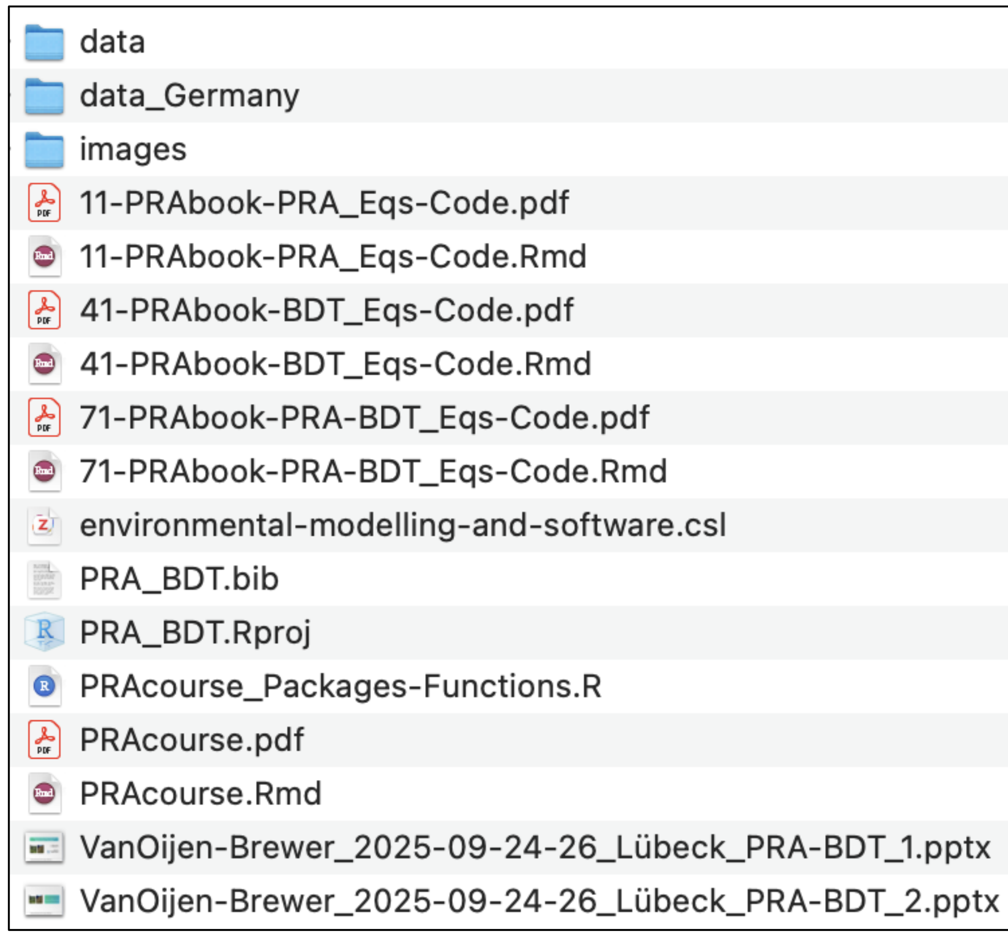
1. **Introduction to PRA: Basic ideas & equations**
2. **Model-based PRA**
3. **Beyond the basic theory**
4. **Introduction to BDT**
5. **Links between PRA and BDT**
6. **General discussion**

- *Contents based on Van Oijen & Brewer (2022), with corrections and extensions*
- *Exercises - simple calculation or R-programming - to initiate discussion*
- *Main questions to all:*
  - ***What is missing from these lectures and/or VO & B (2022)?***
  - ***Which of the methods could you use in your own work?***

Course material

GitHub: https://github.com/MarcelVanOijen/PRA_BDT





References

1.2 References

- Main references: **Van Oijen and Brewer (2022)** ('PRA-BDT') and Van Oijen (2024) chapters 18 and 19 ('BC2').
- Other previous publications: Van Oijen and Zavala (2019), Van Oijen (2019), Van Oijen et al. (2014), Van Oijen et al. (2013).
- Some publications that use (parts of) our PRA approach explicitly: Ren et al. (2023), Nandintsetseg et al. (2021), He et al. (2021), Zhou et al. (2018), Kuhnert et al. (2017), Liu et al. (2020), Nandintsetseg et al. (2024).

1.2.1 Van Oijen and Brewer (2022): Errata

- p.14: the text refers to 'Monte Carlo' (MC) four times but the approach is based on numerical solutions to the conditional expectation values without any MC sampling.
- p.21, 4.3: "any deviation … $E[z \geq thr]$" should be … $E[z|x \geq thr]$.
- pp.13-14,27, Eqs 2.7, 2.8 and 4.5: these are not approximations but exact, so appropriate changes in equations and text must be made.
- p.27: "$Var[z|x < thr] = 1 + \rho\, thr\, E[z] - E[z]^2$" should have conditional expectations on the right: "$... = 1 + \rho\, thr\, E[z|x < thr] - E[z|x < thr]^2$".
- p.27,Eq.4.5: The equations as written give a value equal to 1 for the conditional variances if $\rho = 0$ and $thr \to \infty$ resp. $thr \to -\infty$. The proper limits are σ_z^2 so that is what both appearances of "1" should be replaced by.
- p.59, Eq.11.1: the four R_e should each be Q.

1.
Introduction to PRA: Basic ideas & equations

Terminological confusion

HAZARD	PROBABILITY OF HAZARD	VULNERABILITY
Initiating event	Probability of undesirable consequences	(Severity of) consequences
Threatening event	Exposure	(Degree of) loss
Potentially damaging agent		Biophysical v., Socio-economic v.
Natural disaster		Contextual v., Outcome v.
Natural hazards		Susceptibility (to damage)
		Sensitivity
		Fragility
		Impact
		Damageability

Terminological confusion

Lecina-Diaz et al. (2020):

$$\text{Risk} = E \times HM^S \times LAC \qquad \text{(Equation 1)}$$

"where E are the exposed values (here represented by the ES provided by a forest), HM is the hazard magnitude (weighted by its probability distribution), S is susceptibility, and LAC is the lack of adaptive capacity within a given time frame"

IPCC (2014):

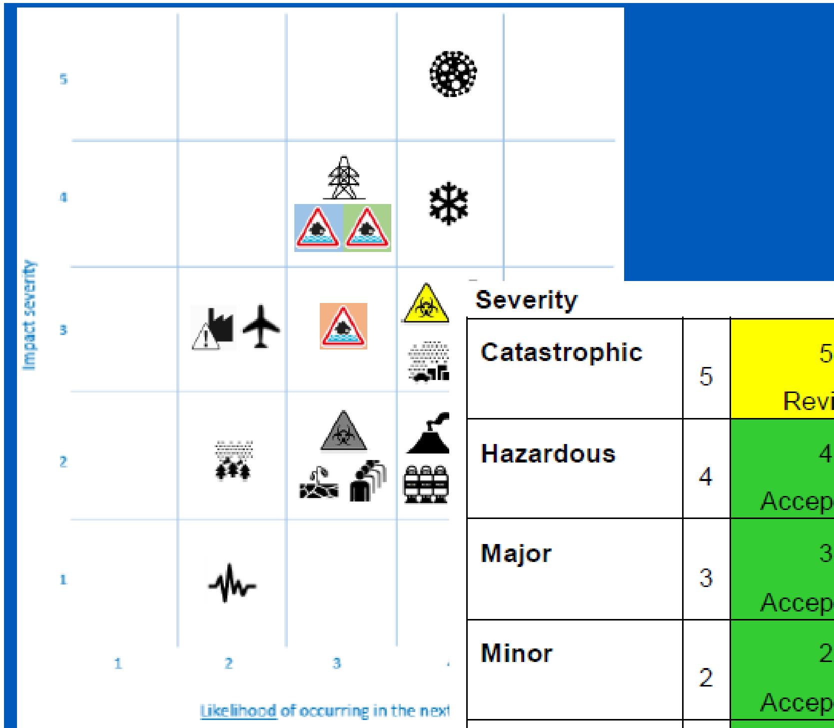
"Risk is often represented as probability of occurrence of hazardous events or trends multiplied by the impacts if these events or trends occur. Risk results from the interaction of vulnerability, exposure, and hazard."

"Vulnerability: The propensity or predisposition to be adversely affected. Vulnerability encompasses a variety of concepts and elements including sensitivity or susceptibility to harm and lack of capacity to cope and adapt."

Risk matrices

National Risk Register (2017)

Matrix A - Hazards, diseases, accidents, and societal risks



CAA (2014)

Severity						
Catastrophic	5	5 Review	10 Unacceptable	15 Unacceptable	20 Unacceptable	25 Unacceptable
Hazardous	4	4 Acceptable	8 Review	12 Unacceptable	16 Unacceptable	20 Unacceptable
Major	3	3 Acceptable	6 Review	9 Review	12 Unacceptable	15 Unacceptable
Minor	2	2 Acceptable	4 Acceptable	6 Review	8 Review	10 Unacceptable
Negligible	1	1 Acceptable	2 Acceptable	3 Acceptable	4 Acceptable	5 Review
		Extremely improbable	**Improbable**	**Remote**	**Occasional**	**Frequent**
		1	2	3	4	5

Likelihood

Risk matrices

$p[H] \times V = 3 \times 4$

$p[H] \times V = 4 \times 3$

$p[H] \times V = 5 \times 2$

What is the highest risk??

V

Severity						
Catastrophic	5	5 Review	10 Unacceptable	15 Unacceptable	20 Unacceptable	25 Unacceptable
Hazardous	4	4 Acceptable	8 Review	12 Unacceptable	16 Unacceptable	20 Unacceptable
Major	3	3 Acceptable	6 Review	9 Review	12 Unacceptable	15 Unacceptable
Minor	2	2 Acceptable	4 Acceptable	6 Review	8 Review	10 Unacceptable
Negligible	1	1 Acceptable	2 Acceptable	3 Acceptable	4 Acceptable	5 Review
		Extremely improbable	**Improbable**	**Remote**	**Occasional**	**Frequent**
		1	2	3	4	5

p[H]

Risk matrices

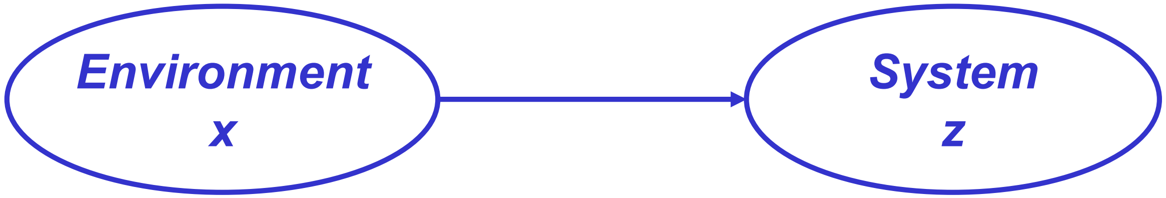
*Can we define V and p[H] such that Risk = V * p[H] is well-defined ?*

V

Severity						
Catastrophic	5	5 Review	10 Unacceptable	15 Unacceptable	20 Unacceptable	25 Unacceptable
Hazardous	4	4 Acceptable	8 Review	12 Unacceptable	16 Unacceptable	20 Unacceptable
Major	3	3 Acceptable	6 Review	9 Review	12 Unacceptable	15 Unacceptable
Minor	2	2 Acceptable	4 Acceptable	6 Review	8 Review	10 Unacceptable
Negligible	1	1 Acceptable	2 Acceptable	3 Acceptable	4 Acceptable	5 Review
		Extremely improbable	**Improbable**	**Remote**	**Occasional**	**Frequent**
		1	2	3	4	5

p[H]

Toward formal theory for PRA: Notation



H	= Hazardous conditions
$p[x \in H]$	= p[H]
	= Probability of hazardous conditions
$E[z]$	= Expectation for system performance
$E[z \mid H]$	= Expectation for system performance under hazardous (”bad”) conditions
$E[z \mid \neg H]$	= Expectation for system performance under non-hazardous (“good”) conditions

PRA: basic equations

$$R = E[z|\neg H] - E[z]$$

Standard definition of *risk*

$$p[H] = p[x \in H]$$

Standard definition of *hazard probability*

$$V = R \,/\, p[H]$$

Logical definition of *vulnerability*

$$\begin{aligned} &= (E[z|\neg H] - E[z]) \,/\, p[H] \\ &= (E[z|\neg H] - p[H]E[z|H] - (1 - p[H])E[z|\neg H]) \,/\, p[H] \\ &= E[z|\neg H] - E[z|H]. \end{aligned}$$

Vulnerability (V) is the difference in expected system performance (E[z|.]) between non-hazardous (¬H) and hazardous conditions (H).

Data needs are minimal for basic PRA

Cause **Effect**

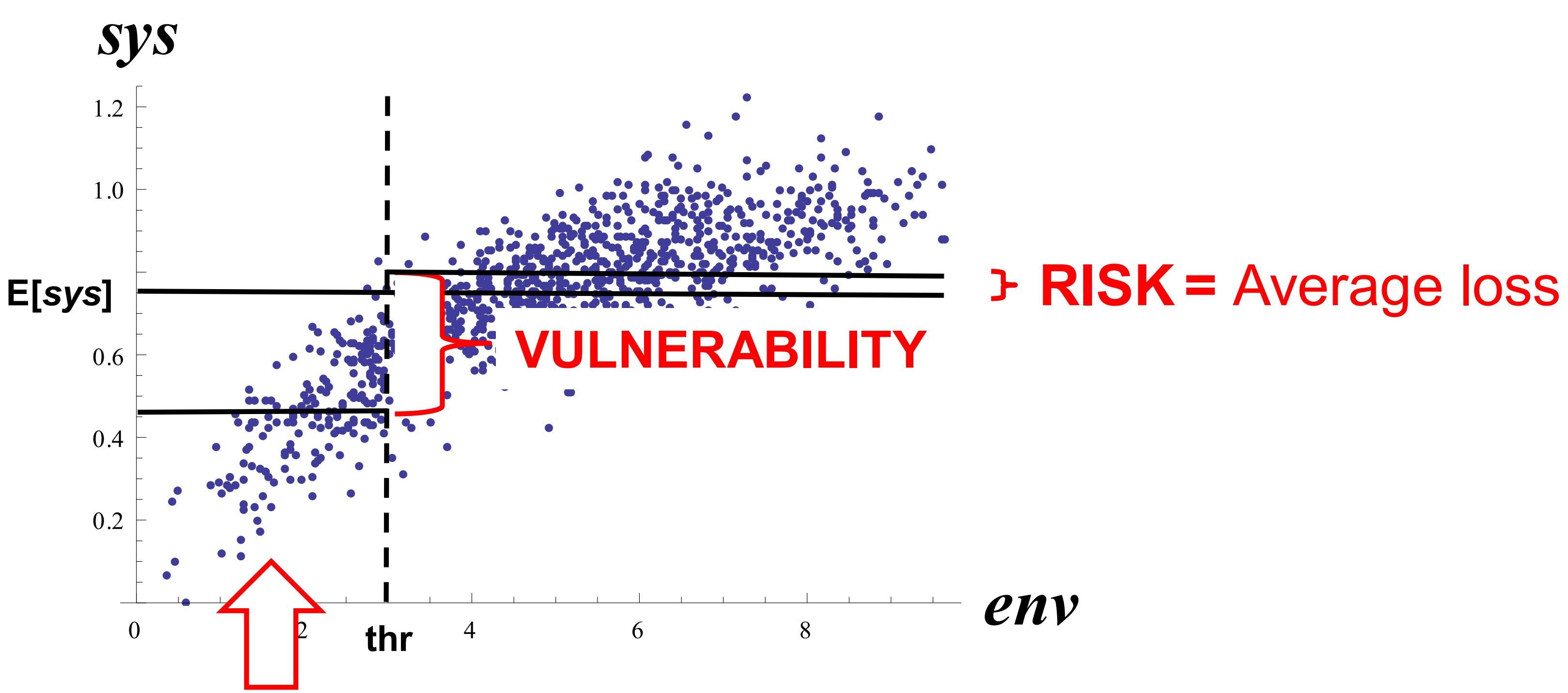
env (rainfall, mm d^{-1})	*sys* (NPP, g C m^{-2} d^{-1})
1	**2**
2	**3**
2	**4**
3	7
3	8
3	8
4	9
4	10
4	10
4	11

"Hazardous conditions" = "rainfall <= 2 mm d^{-1}"

A: E(*sys*|*env* non-hazardous) = 9 g m^{-2} d^{-1}
B: E(*sys*|*env* hazardous) = 3 g m^{-2} d^{-1}
C: Vulnerability = A-B = 6 g m^{-2} d^{-1}
D: p[*env* hazardous] = 0.3
E: Risk = C*D = 1.8 g m^{-2} d^{-1}

Note that Risk can also be calculated in the standard way as E[sys|env non-hazardous] – E[sys] = 9 - 7.2 = 1.8

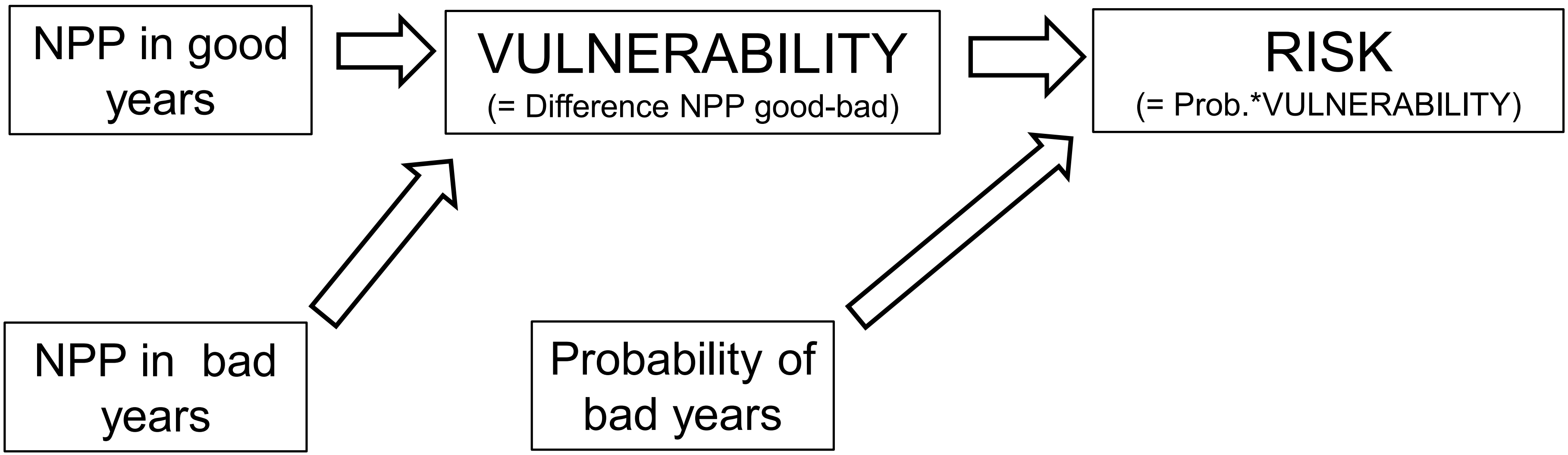
A simple method for risk decomposition



p[Hazardous] =
frequency *env* < thr =
n_H / n

RISK =
VULNERABILITY * p[Hazardous]

Example: Risk analysis for NPP



PRA coniferous forests (WATCH, EMEP, BASFOR)

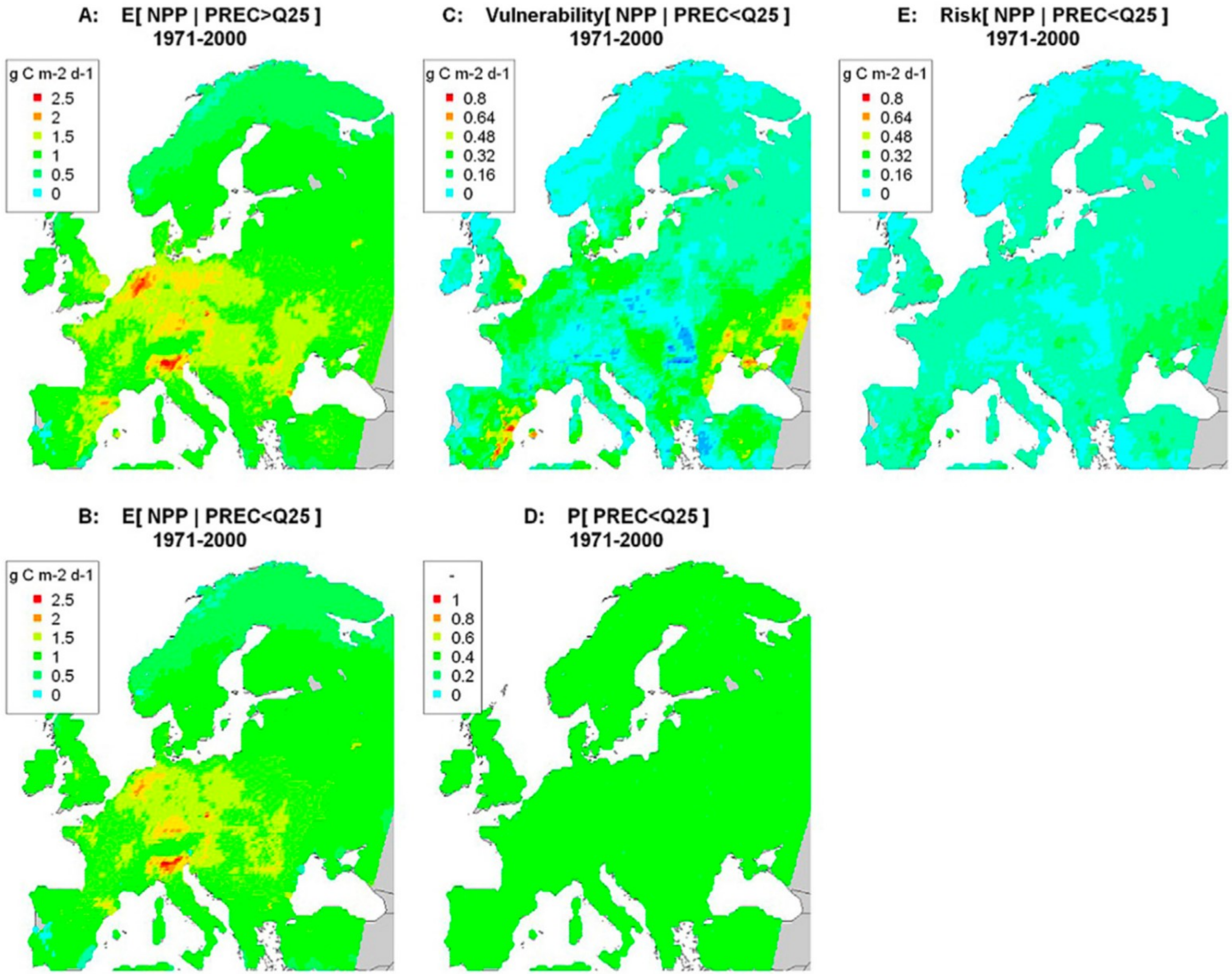


Figure 1. Risk analysis for NPP in 1971–2000 as affected by precipitation. 'Hazardous' years have less precipitation than the 25% quantile (Q25). (A): NPP in hazardous years; (B): NPP in non-hazardous years; (C): Vulnerability $= A - B$; (D): Probability of hazardous years; (E): Risk $= C * D$. Note that in this example, risk (E) is a rescaling of vulnerability (C) because of the constant probability of hazardousness (D).

PRA coniferous forests (MPI-Remo, EMEP, BASFOR)

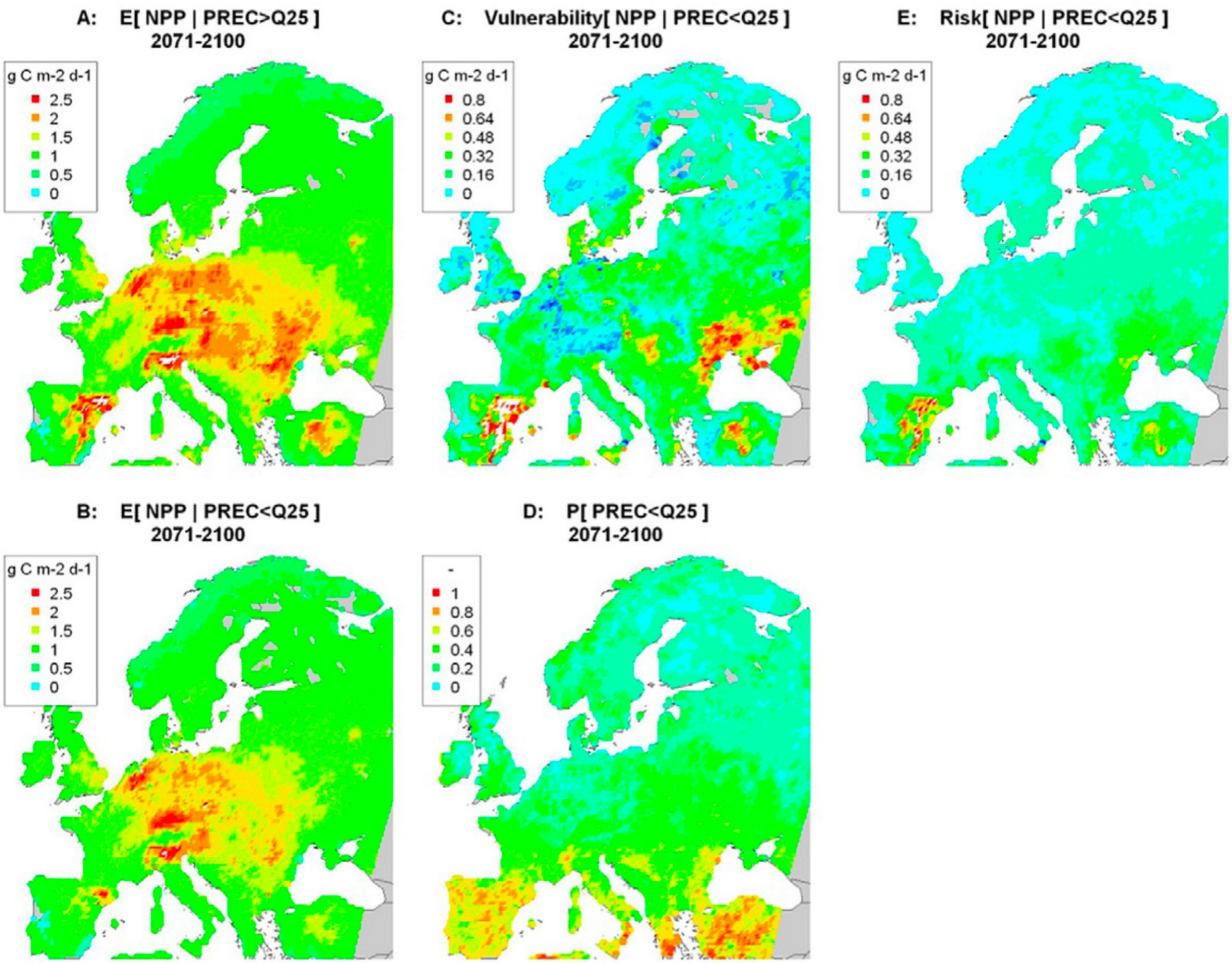
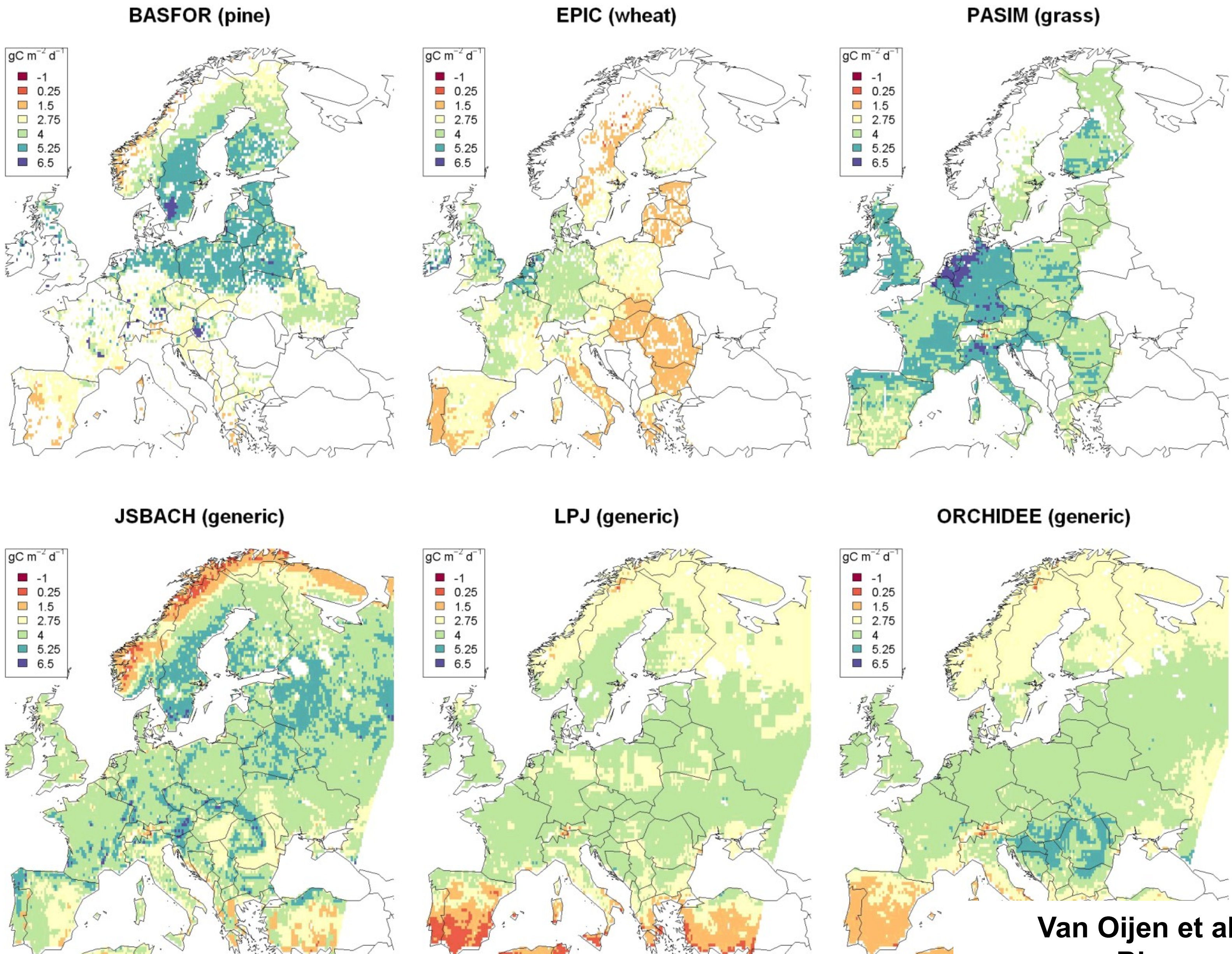


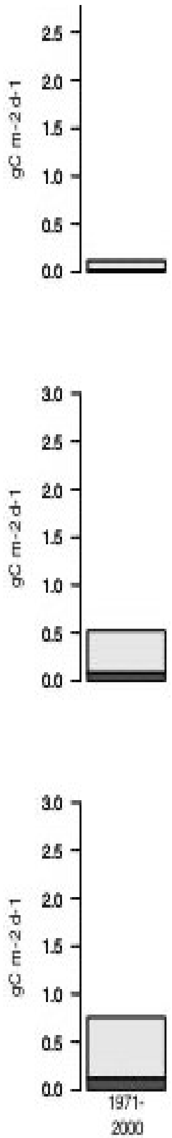
Figure 2. As figure 1 for 2071–2100. Note that 'hazardousness' is unchanged: Q25 still refers to 1971–2000.

6 models: average NPP in 1971-2000



Van Oijen et al. (2014) Biogeosciences

Risk analysis NPP (model LPJ) per latitudinal band



VULNERABILITY (high-latitude)

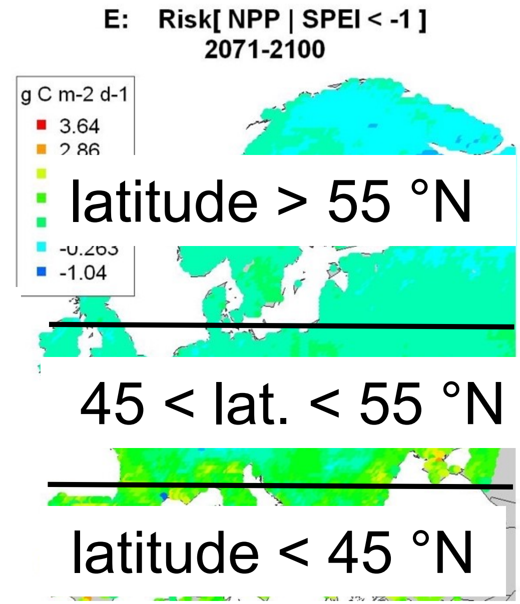
RISK (high-latitude)

VULNERABILITY (mid-latitude)

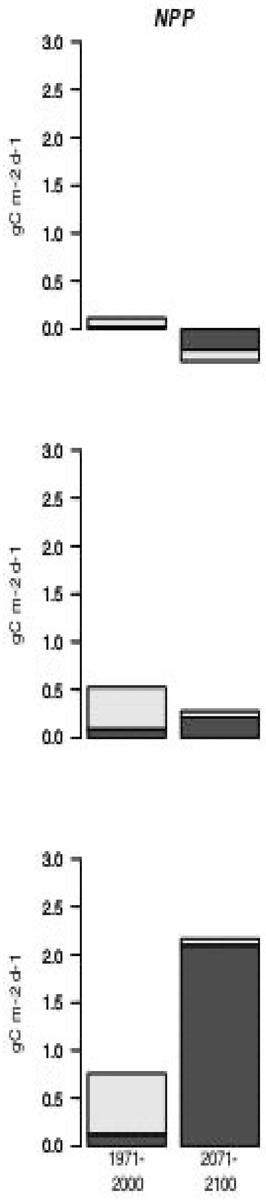
RISK (mid-latitude)

VULNERABILITY (low-latitude)

RISK (low-latitude)

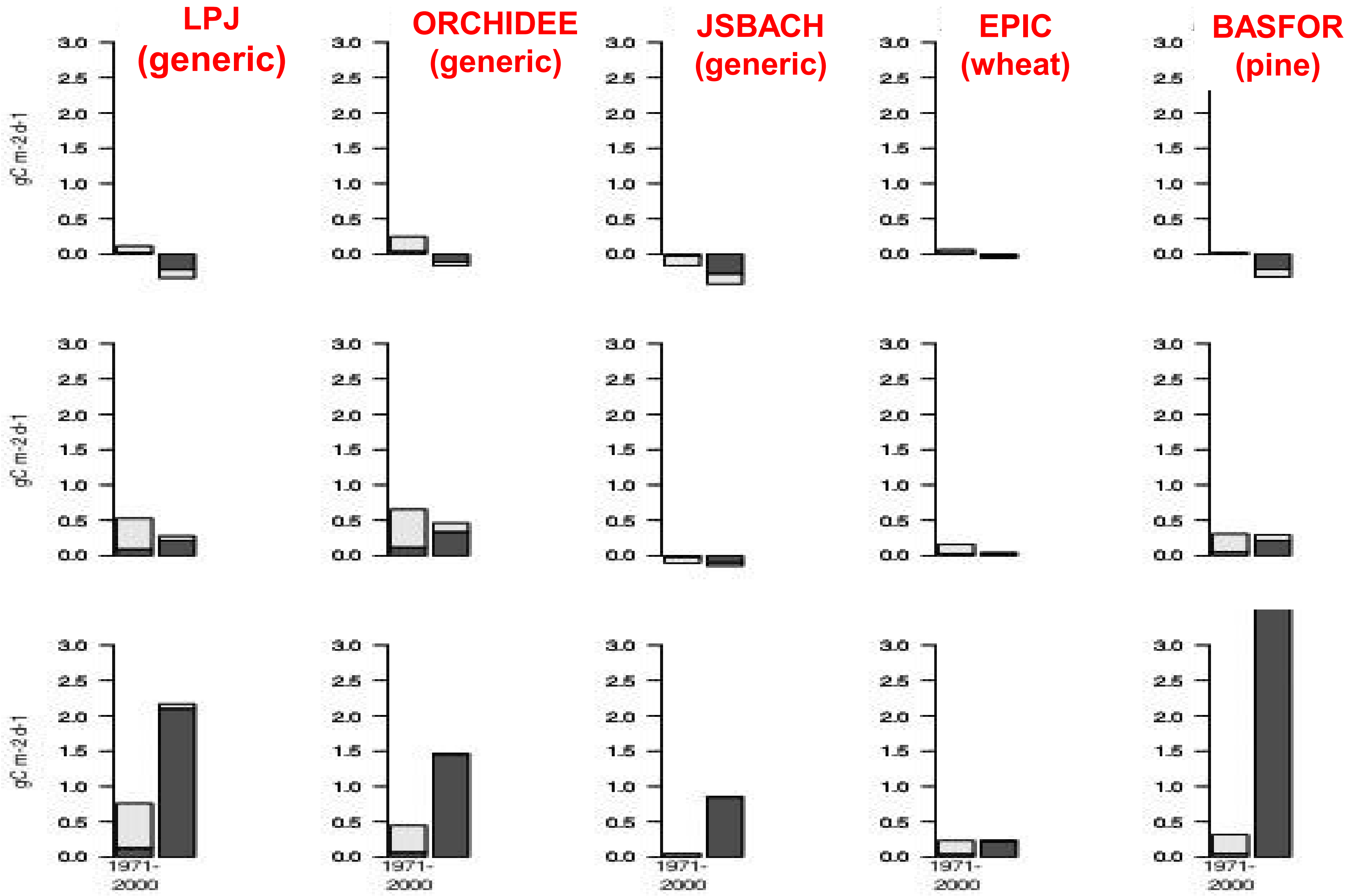


Risk analysis NPP (model LPJ) per latitudinal band



Left: 1971-2000, Right: 2071-2100

Risk analysis NPP: 5 models



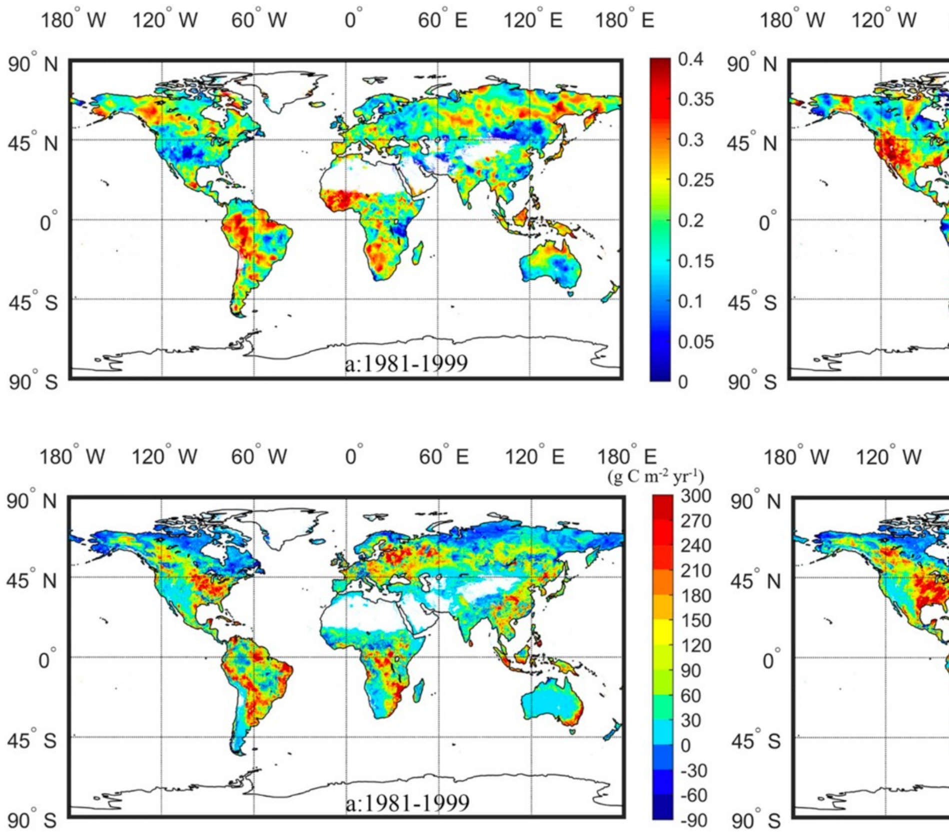
He et al. (2021). Drought Risk of Global Terrestrial Gross Primary Productivity Over the Last 40 Years Detected by a Remote Sensing-Driven Process Model

Journal of Geophysical Research: Biogeosciences 10.1029/2020JG005944

1981-1999 **2000-2016**

p[H]

V



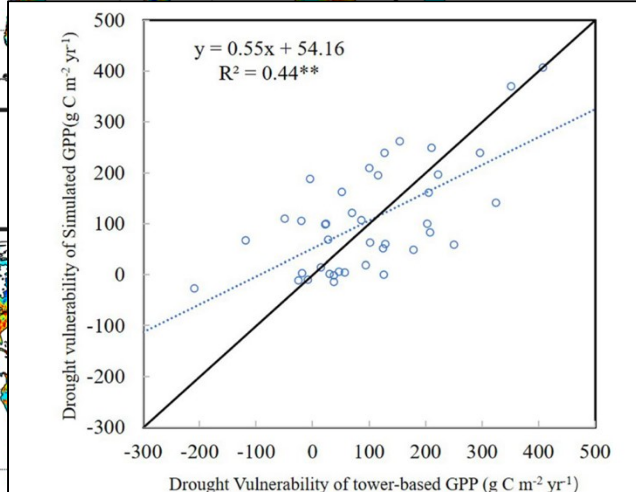


Figure 2. Comparison of drought vulnerability of simulated gross primary productivity (GPP) against that of tower-based GPP at 41 eddy-covariance sites. The solid black line is the 1:1 line and dashed dark blue line is the regression line.

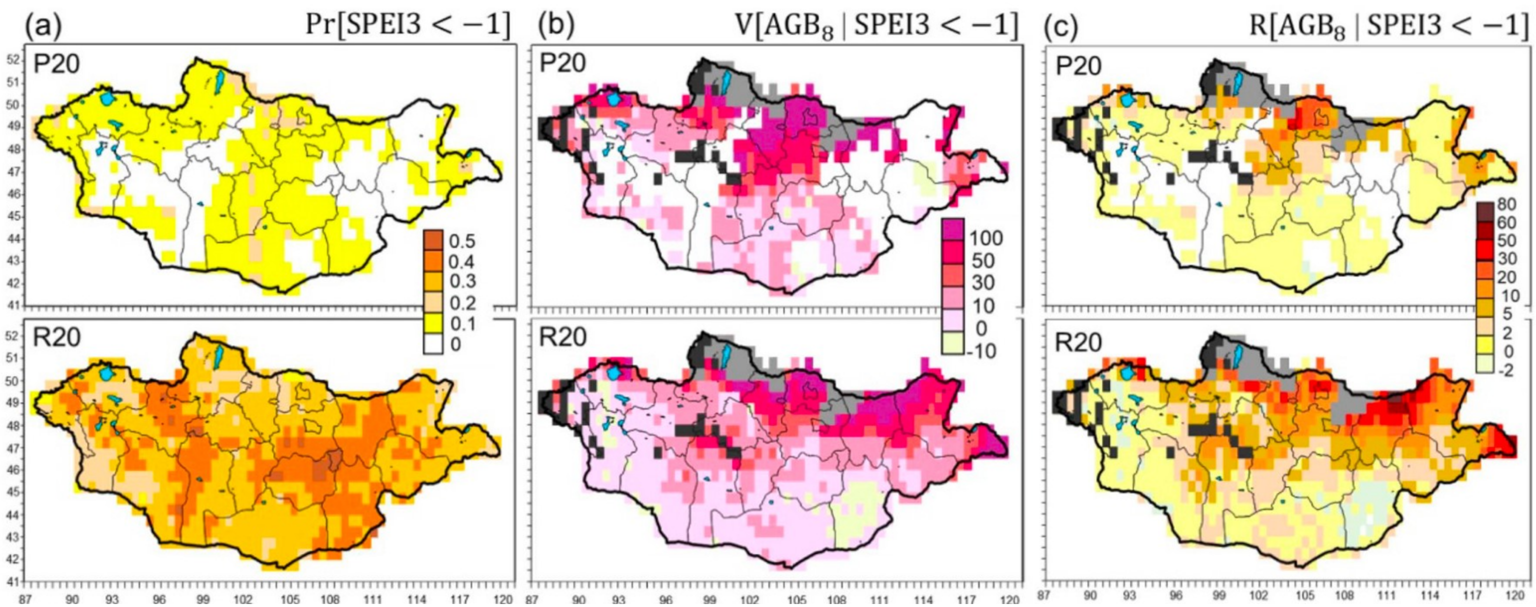
Nandintseg et al. (2021). Risk and vulnerability of Mongolian grasslands under climate change

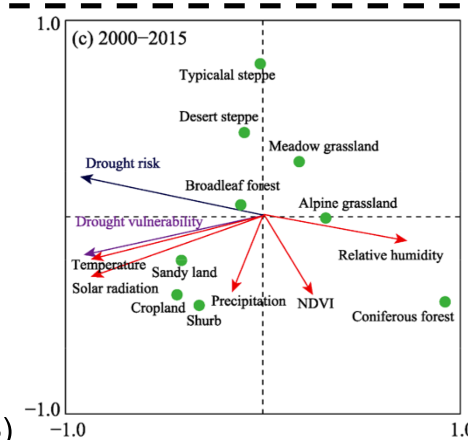
Environ. Res. Lett. **16** (2021) 034035

p[H] V R

1976-1995

1996-2015



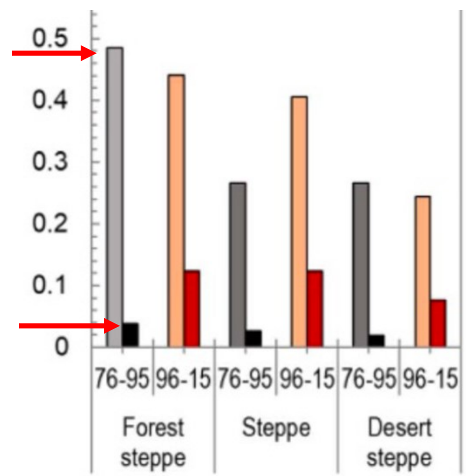


Ren et al. (2023)

Figure 10 Redundancy analysis of the drought risk and vulnerability of net primary productivity (NPP)

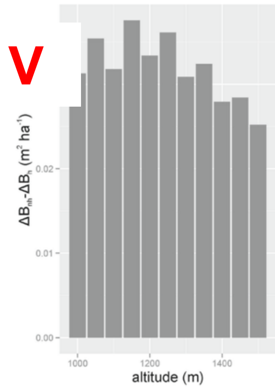
V

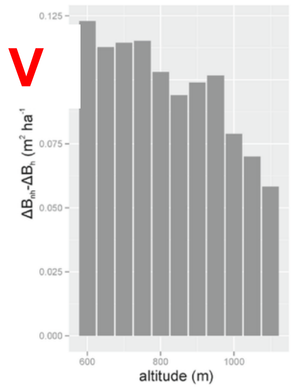
R

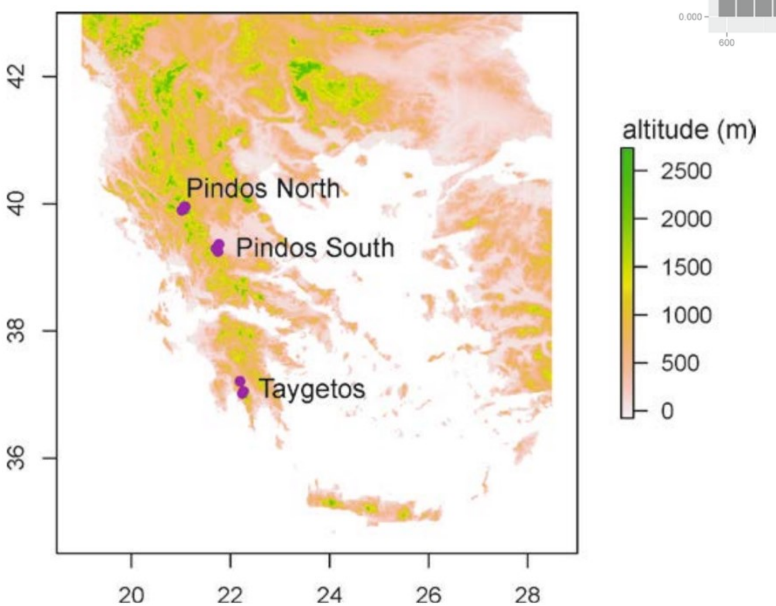


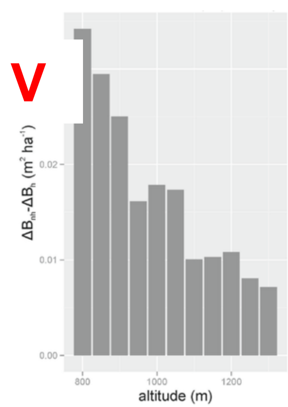
Fyllas et al. (2015). Mediterranean Forests in Transition (MEDIT): Deliverable No9

Vulnerability of tree growth to climate change as f(altitude)









Zhou et al. (2018). Drought Impacts on Vegetation Indices and Productivity of Terrestrial Ecosystems in Southwestern China During 2001–2012

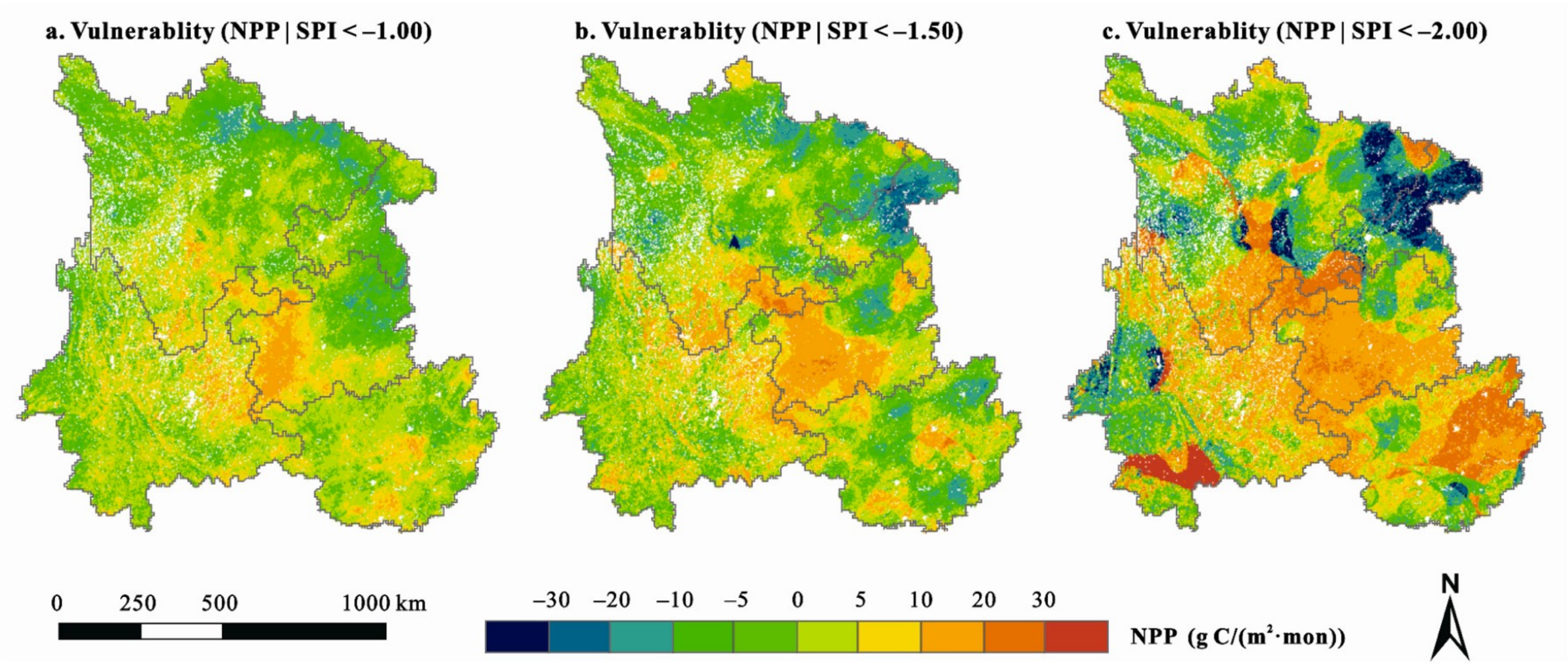


Fig. 7 Vulnerability analysis for the Net Primary Productivity (NPP) in 2001–2010 using CASA model. a, b and c represented the vulnerability of NPP to drought when SPI was lower than –1.00, –1.50 and –2.00, respectively

Kuhnert et al. (2017). Impact analysis of climate data aggregation at different spatial scales on simulated net primary productivity for croplands

Drought risk to NPP of maize in North Rhine-Westphalia

V, R

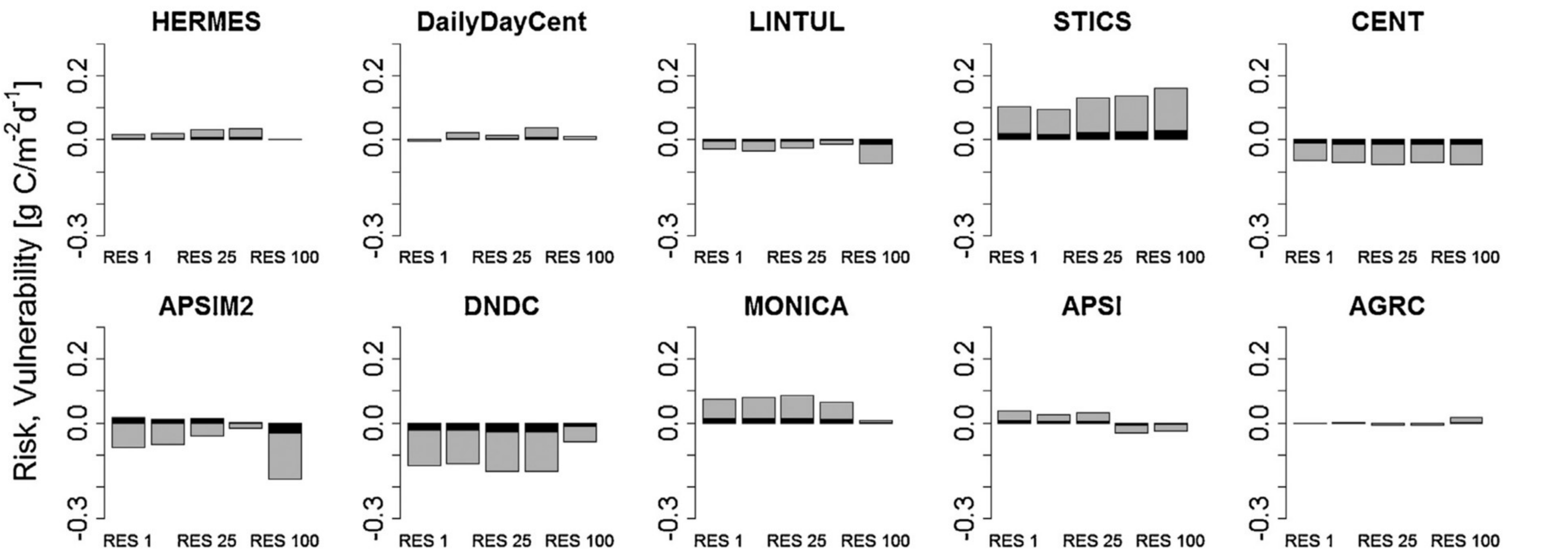
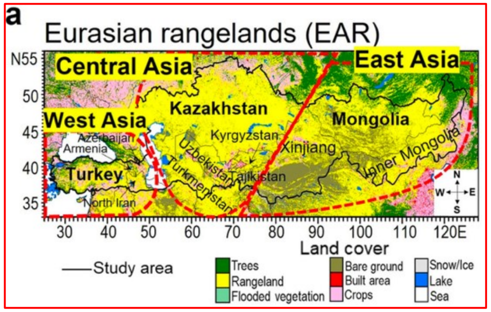
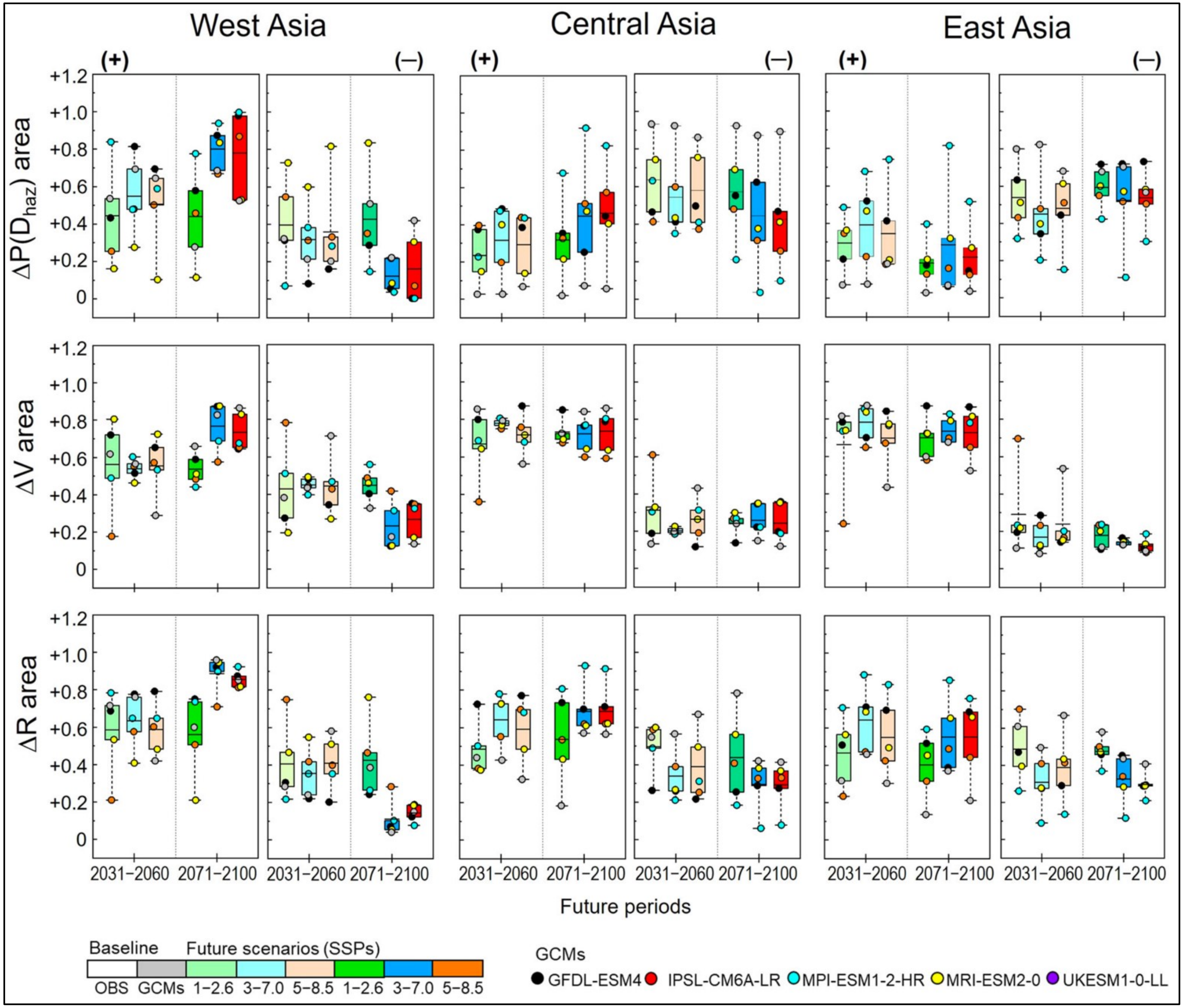


Fig. 9. Vulnerability (grey bar) and risk (black bar) for 10 models simulated for water limitation considered (WN). The results represent the simulation results for maize for the period 1983–2011. The terms vulnerability and risk are used according to the definition by van Oijen et al. (2014) and describe the impacts of hazardous in comparison to non-hazardous conditions (see also section 2.6.

Nandintseg et al. (2024). Future drought risk and adaptation of pastoralism in Eurasian rangelands

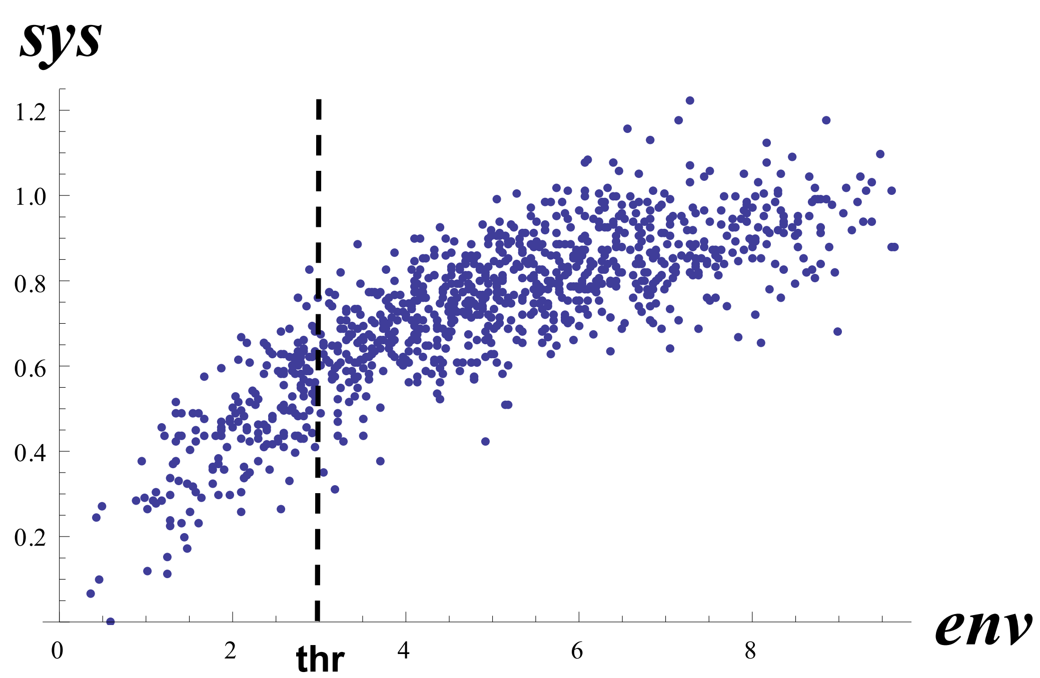




Uncertainty p[H]

Uncertainty p[H]

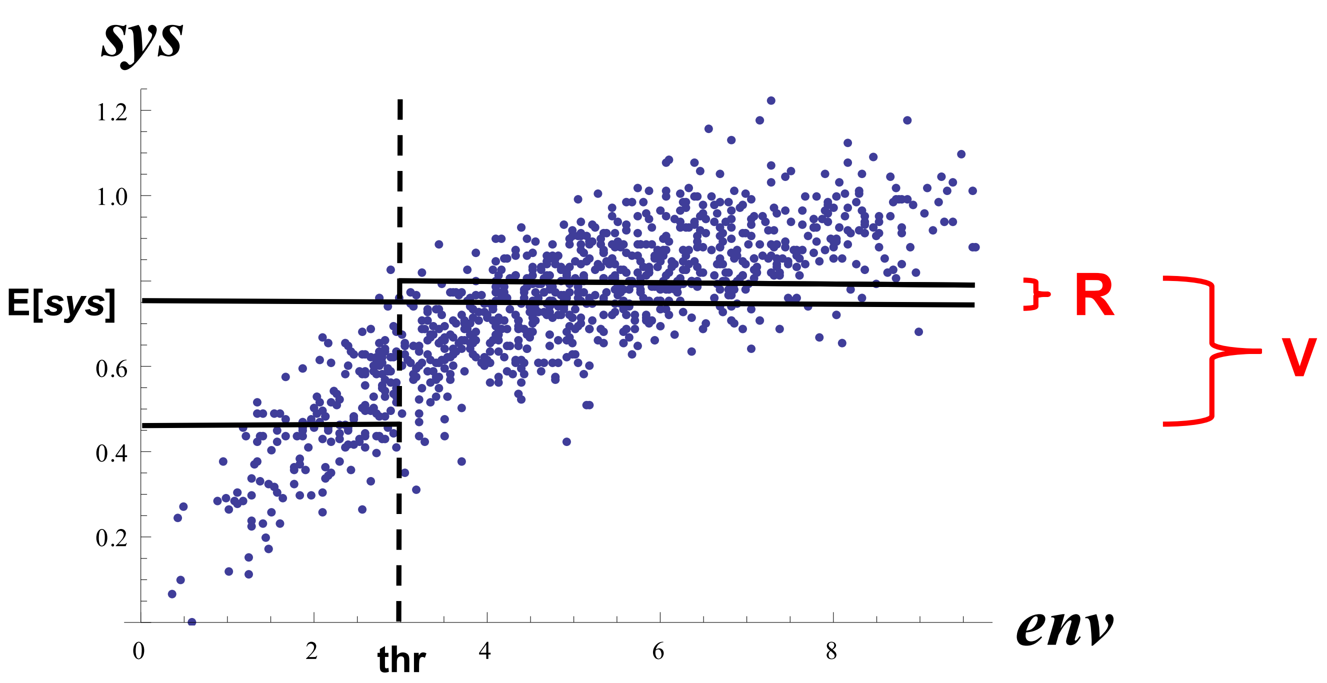
Uncertainty Quantification (UQ)



n_H
observations **below** threshold

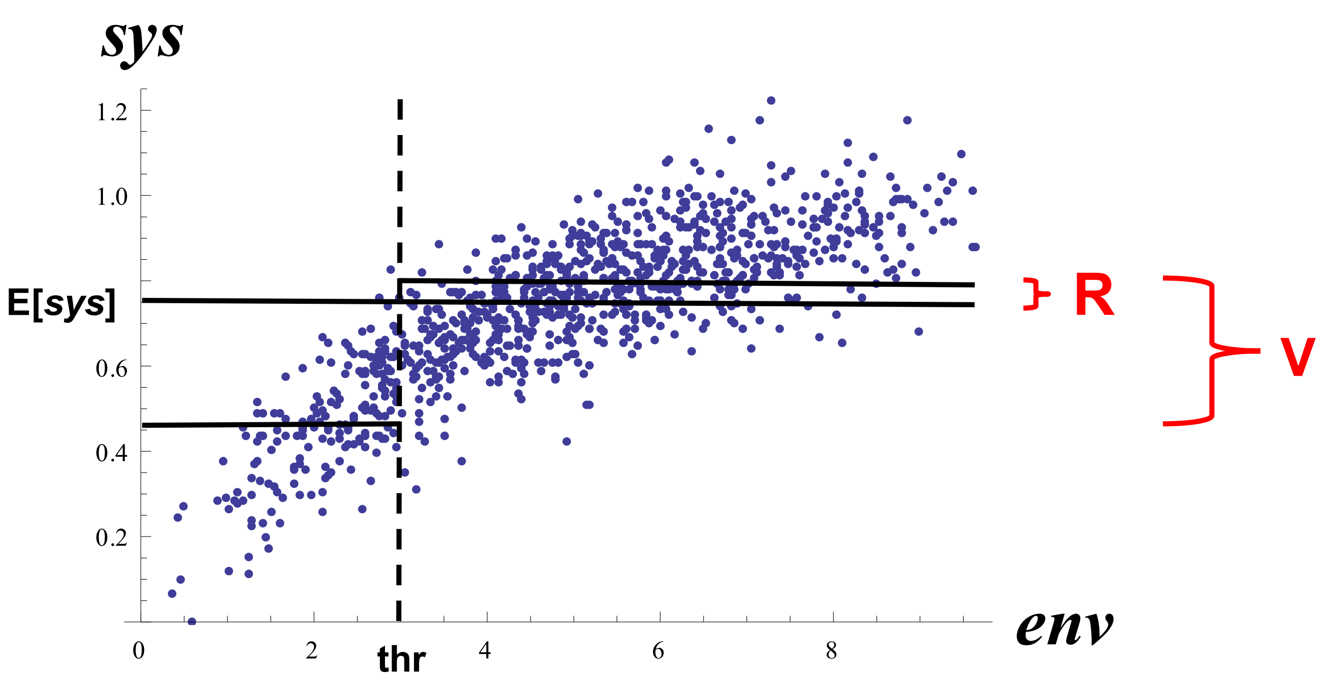
$n\text{-}n_H$
observations **above** threshold

Uncertainty Quantification (UQ)



$p[H] = n_H / n$

Uncertainty Quantification (UQ)



p[H] = n_H / n

$$\sigma^2_{p[H]} = p[H]\,(1 - p[H])/n$$

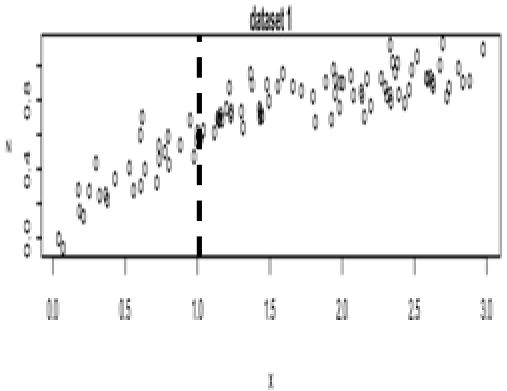
$$\sigma^2_V = \frac{Var[z_H]}{n_H} + \frac{Var[z_{\neg H}]}{n - n_H}$$

$$\sigma^2_R = \sigma^2_{p[H]}\sigma^2_V + \sigma^2_{p[H]}V + p[H]\sigma^2_V$$

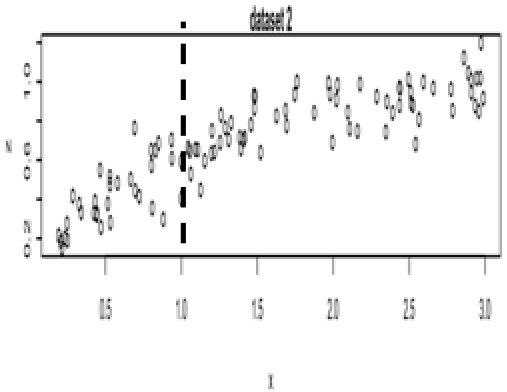
R-function for single-threshold sampling-based PRA

```
PRA <- function( x, z, thr=0 ) {
  n     <- length(x)
  H     <- which(x < thr) ; n_H <- length(H)
  Ez_H <- mean( z[ H] )   ; Ez_NotH   <- mean( z[-H] )

  pH    <- n_H / n
  V     <- Ez_NotH - Ez_H
  R     <- pH * V

  s_pH <- sqrt( pH*(1-pH) / n )
    s_Ez_H     <- sqrt( var(z[ H]) /     n_H   )
    s_Ez_NotH <- sqrt( var(z[-H]) / (n-n_H) )
  s_V   <- sqrt( s_Ez_H^2 + s_Ez_NotH^2 )
  s_R   <- sqrt( s_pH^2*s_V^2 + s_pH^2*V^2 + pH^2*s_V^2 )

  return( c(pH=pH, V=V, R=R, s_pH=s_pH, s_V=s_V, s_R=s_R) )
}
```

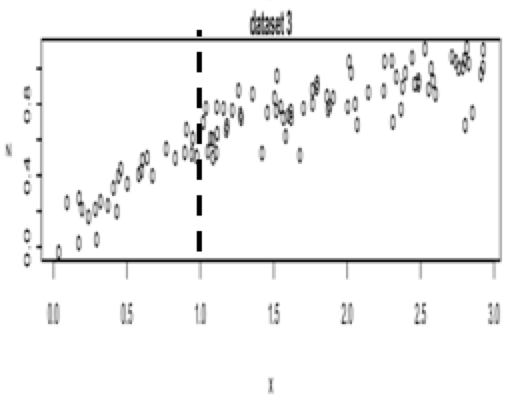
UQ: 1000 datasets from same distribution, thr=1



p[H] = 0.27, V = 0.464, R = 0.124
$\boldsymbol{\sigma}_{pH}$ = 0.04 , $\boldsymbol{\sigma}_{V}$ = 0.040, $\boldsymbol{\sigma}_{R}$ = 0.018

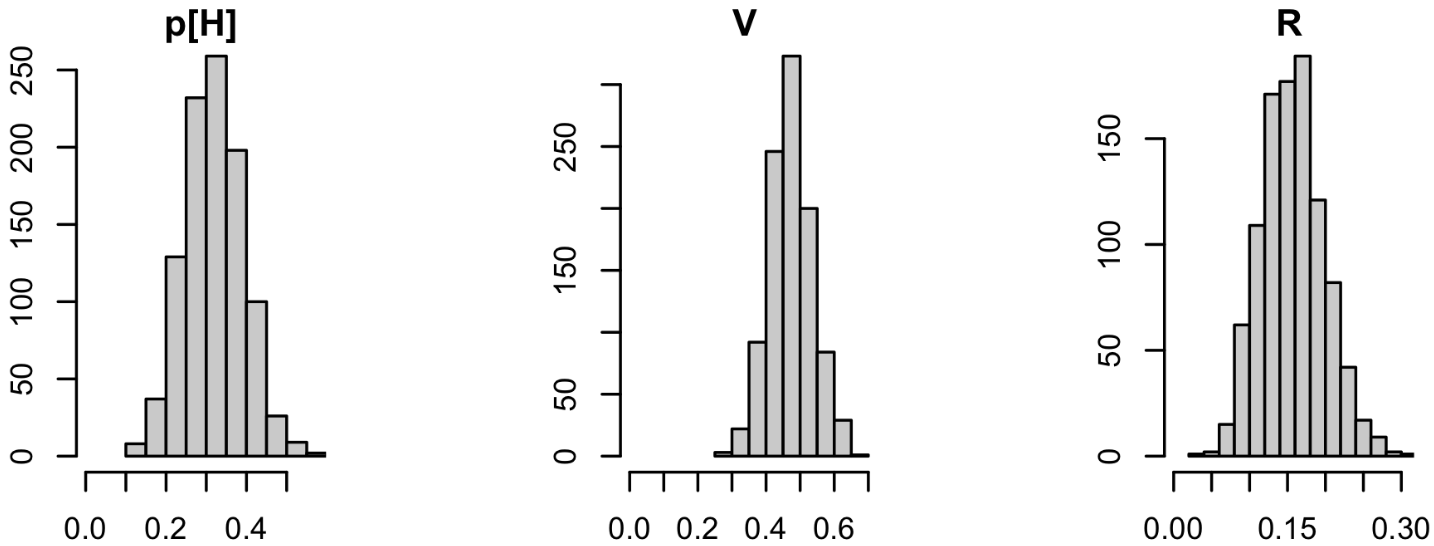


p[H] = 0.33, V = 0.410, R = 0.135
$\boldsymbol{\sigma}_{pH}$ = 0.05. , $\boldsymbol{\sigma}_{V}$ = 0.035, $\boldsymbol{\sigma}_{R}$ = 0.019

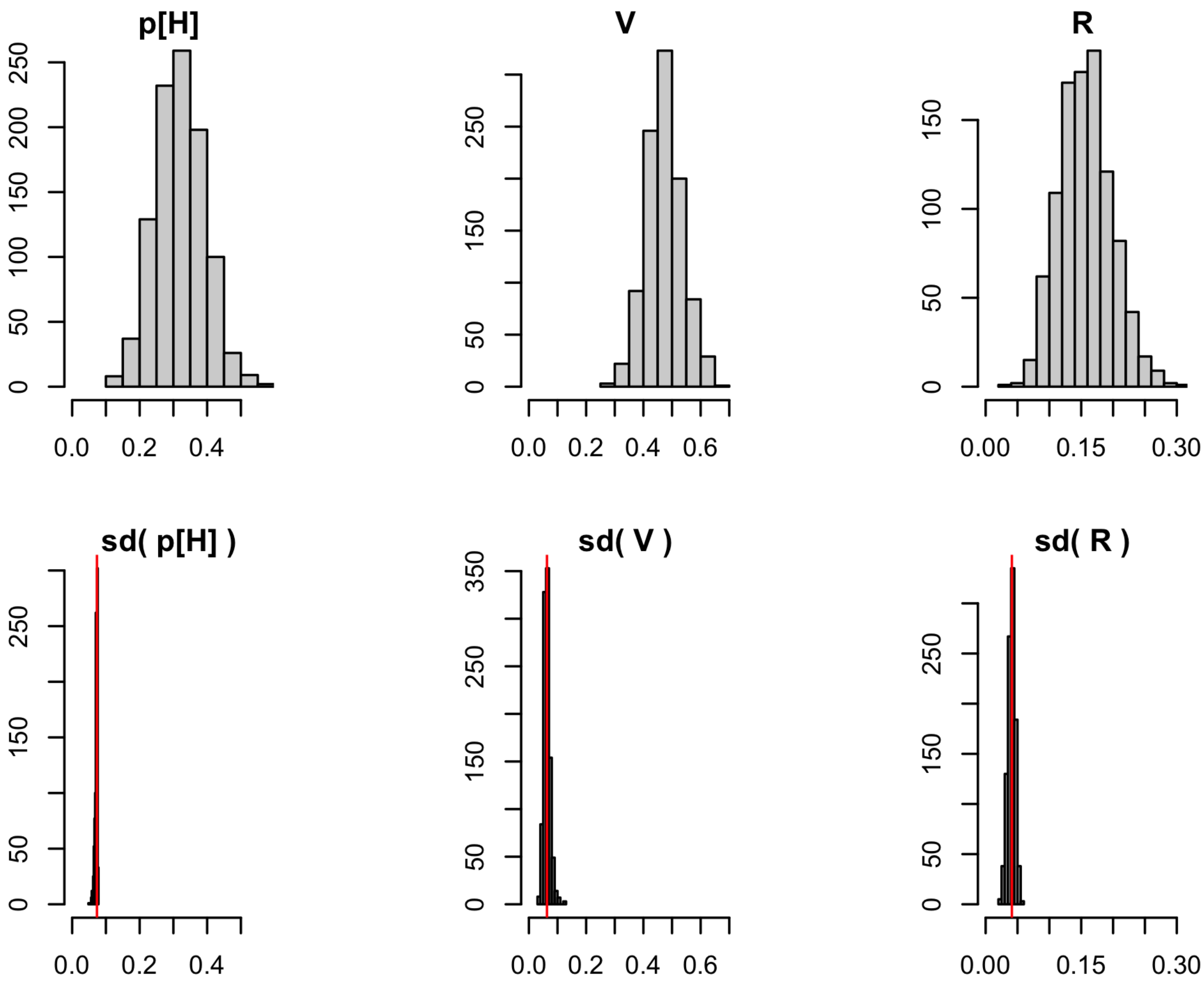


p[H] = 0.27, V = 0.464, R = 0.124
$\boldsymbol{\sigma}_{pH}$ = 0.04 , $\boldsymbol{\sigma}_{V}$ = 0.039, $\boldsymbol{\sigma}_{R}$ = 0.019

UQ: 1000 datasets from same distribution, thr=1



UQ: 1000 datasets from same distribution, thr=1



EXERCISE. Minimal data sets (sampling-based PRA)

```
x <- c( 200, 600 )
z <- c(  70,  90 )
PRA( x, z, 500 )
```

Will the uncertainty quantification work?

```
x <- c( 200, 400, 600, 800 )
z <- c(  70,  80,  90, 100 )
PRA( x, z, 500 )
```

Will the uncertainty quantification work?

EXERCISE. Minimal data sets (sampling-based PRA)

```
x <- c( 200, 600 )
z <- c(  70,  90 )
PRA( x, z, 500 )
```

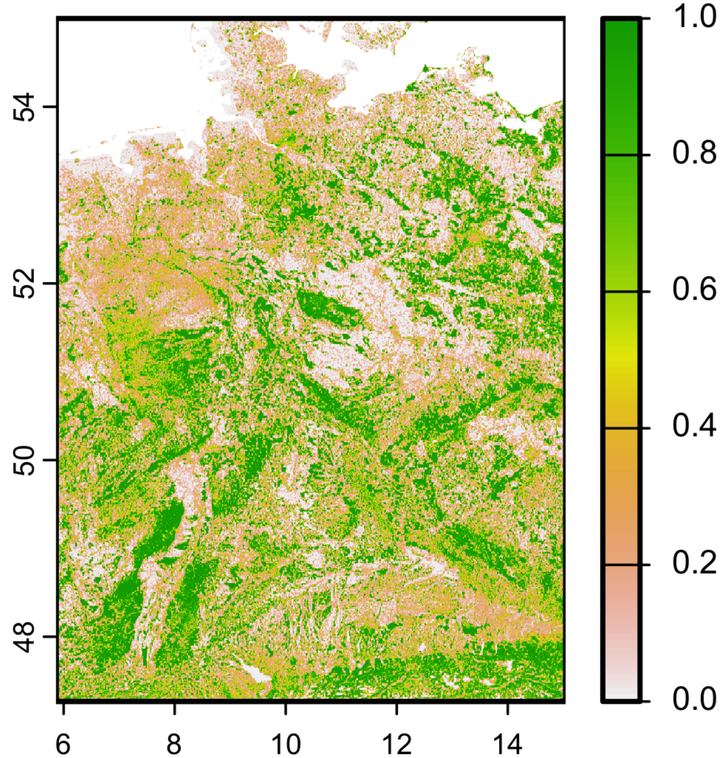
```
pH   V    R    s_pH      s_V       s_R
0.5  20.0 10.0 0.3535534 NA        NA
```

```
x <- c( 200, 400, 600, 800 )
z <- c(  70,  80,  90, 100 )
PRA( x, z, 500 )
```

```
pH   V    R    s_pH      s_V       s_R
0.5  20.0 10.0 0.25      7.0710678 6.373774
```

Tree cover Germany 2020 (ESA WorldCover)

```
r_tree_DEU <- rast( "data/landuse/r_tree_DEU.tif" )
plot( r_tree_DEU, col=terrain.colors(100,rev=T) )
```



Forest data from Germany

https://www.umweltbundesamt.de/daten/klima/

- 2-bis-7_abb-tab_nsh_2024-04-08.xlsx

https://wo-apps.thuenen.de/apps/wze/

- absterberate_EI_GFI_GKI_RBU_zeitreihe.csv

```
library( readxl )

d.data         <- "data_Germany/"

file.data      <- paste0( d.data, "2-bis-7_abb-tab_nsh_2024-04-08.xlsx" )
sheet.data     <- "5_Daten" ; r.data <- "B120:C153"
y_r            <- read_excel( file.data, sheet=sheet.data, range=r.data,
                              col_names=F )
y_r            <- round( as.matrix(y_r) )
colnames(y_r) <- c( "year", "rain.mm" )
```

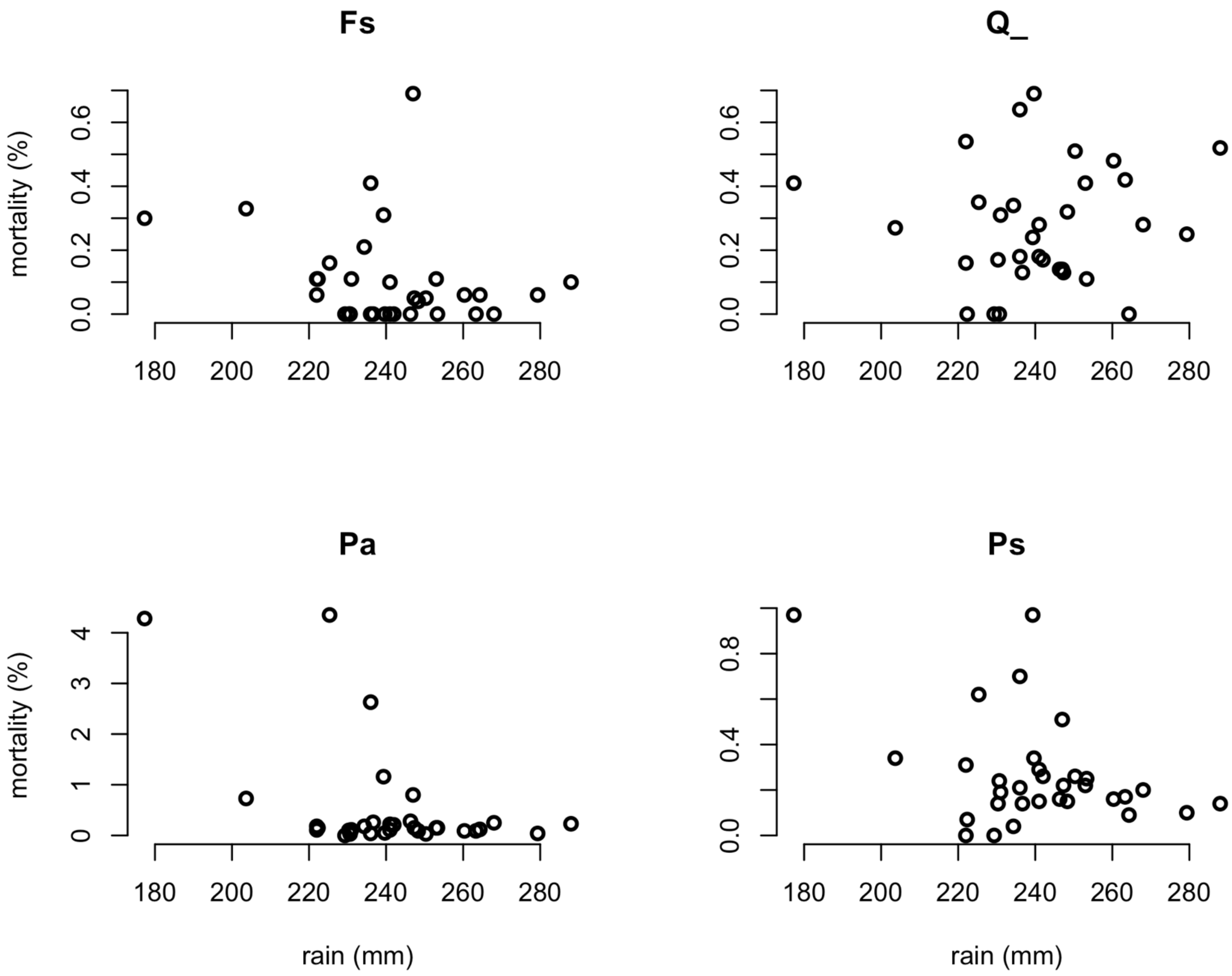
```
file.data      <- paste0( d.data, "absterberate_EI_GFI_GKI_RBU_zeitreihe.csv" )
y_m            <- read.csv2( file.data )
colnames(y_m) <- c( "year", "mort.%.Fs", "mort.%.Q_", "mort.%.Pa", "mort.%.Ps" )
y_r_m          <- cbind( y_r, y_m[ , startsWith( colnames(y_m), "mort" ) ] )
```

Forest data from Germany

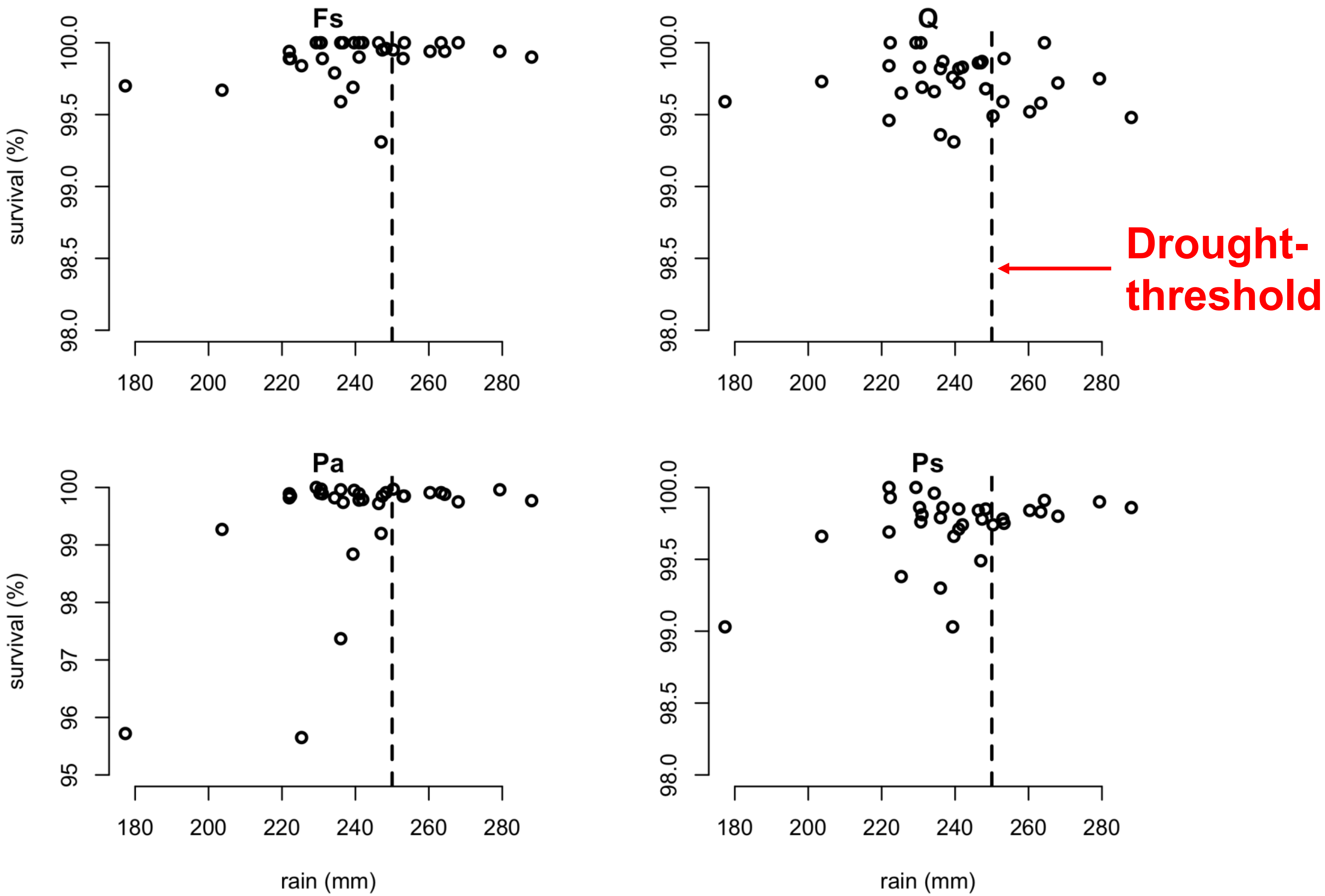
```
# Annual mortality vs. summer rain of last three years
  meanlast <- function(v,k){as.vector(filter(v,f=rep(1/k,k),s=1))}
  n   <- dim(y_r_m)[1] ; k <- 3
  r.3 <- meanlast(y_r_m[,"rain.mm"],k)[k:n]
```

```
  z.Fs <- y_r_m[ k:n, "mort.%.Fs” ]
  z.Q_ <- y_r_m[ k:n, "mort.%.Q_” ]
  z.Pa <- y_r_m[ k:n, "mort.%.Pa” ]
  z.Ps <- y_r_m[ k:n, "mort.%.Ps” ]

  par ( mfrow=c(2,2), mar=c(5,4,4,2) )
  plot( r.3, z.Fs, main="Fs", xlab=""          , ylab="mortality (%)" )
  plot( r.3, z.Q_, main="Q_", xlab=""          , ylab=""              )
  plot( r.3, z.Pa, main="Pa", xlab="rain (mm)", ylab="mortality (%)" )
  plot( r.3, z.Ps, main="Ps", xlab="rain (mm)", ylab=""              )
```

Forest mortality data from Germany



Forest survival data from Germany

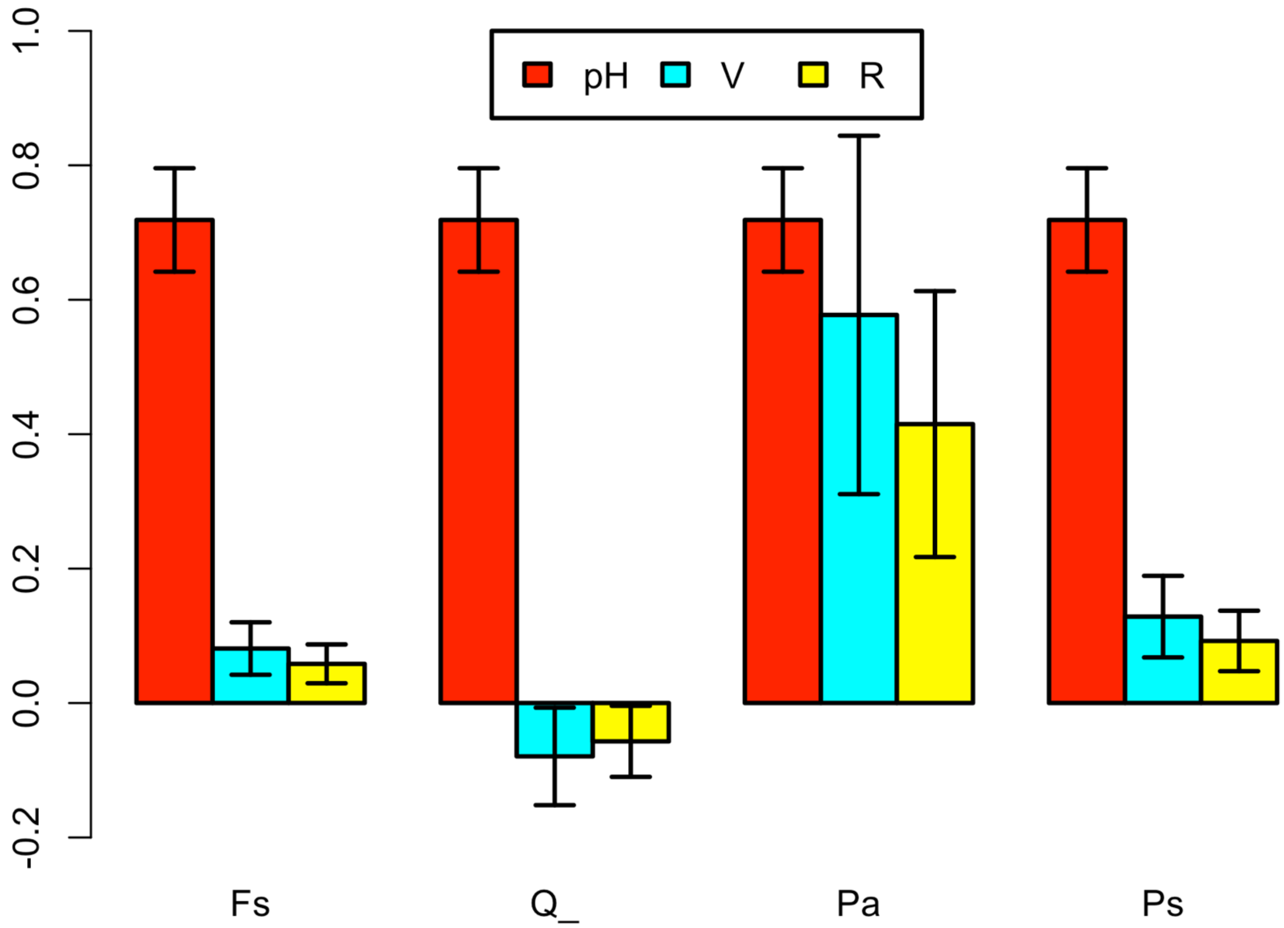


Forest data from Germany: PRA

```
# PRA for survival (= 100 - mortality)
  PRA_Ss.Fs <- PRA( x_r3, 100-z_Fs, thr=250 )
  PRA_Ss.Q  <- PRA( x_r3, 100-z_Q , thr=250 )
  PRA_Ss.Pa <- PRA( x_r3, 100-z_Pa, thr=250 )
  PRA_Ss.Ps <- PRA( x_r3, 100-z_Ps, thr=250 )
```

```
PRA._ <- as.matrix( cbind(PRA_Ss.Fs, PRA _Ss.Q_, PRA _Ss.Pa, PRA _Ss.Ps) )
colnames(PRA._) <- c( "Fs", "Q_", "Pa", "Ps" )
means <- PRA._[1:3,] ; s <- PRA._[4:6,]

par(mfrow=c(1,1))
col   <- c( "red", "cyan", "yellow" )
bp    <- barplot( means, beside=T, ylim=c(-0.2,1), col=col,
                  legend.text=row.names(means),
                  args.legend=list(x="top",hor=T) )
segments( bp, means - s, bp, means + s )
ew    <- (bp[2,1]-bp[1,1])/4
segments( bp - ew, means - s, bp + ew, means - s )
segments( bp - ew, means + s, bp + ew, means + s )
```

Forest data from Germany: PRA



Forest data from Germany: EXERCISE 1

Discuss the limitations of this PRA.

Forest data from Germany: EXERCISE 2

Our original R-code for PRA on *Picea abies* was:

- `PRA(  x_r3, 100-z_Pa,  250 )`

Run the following lines instead. Compare with original and explain.

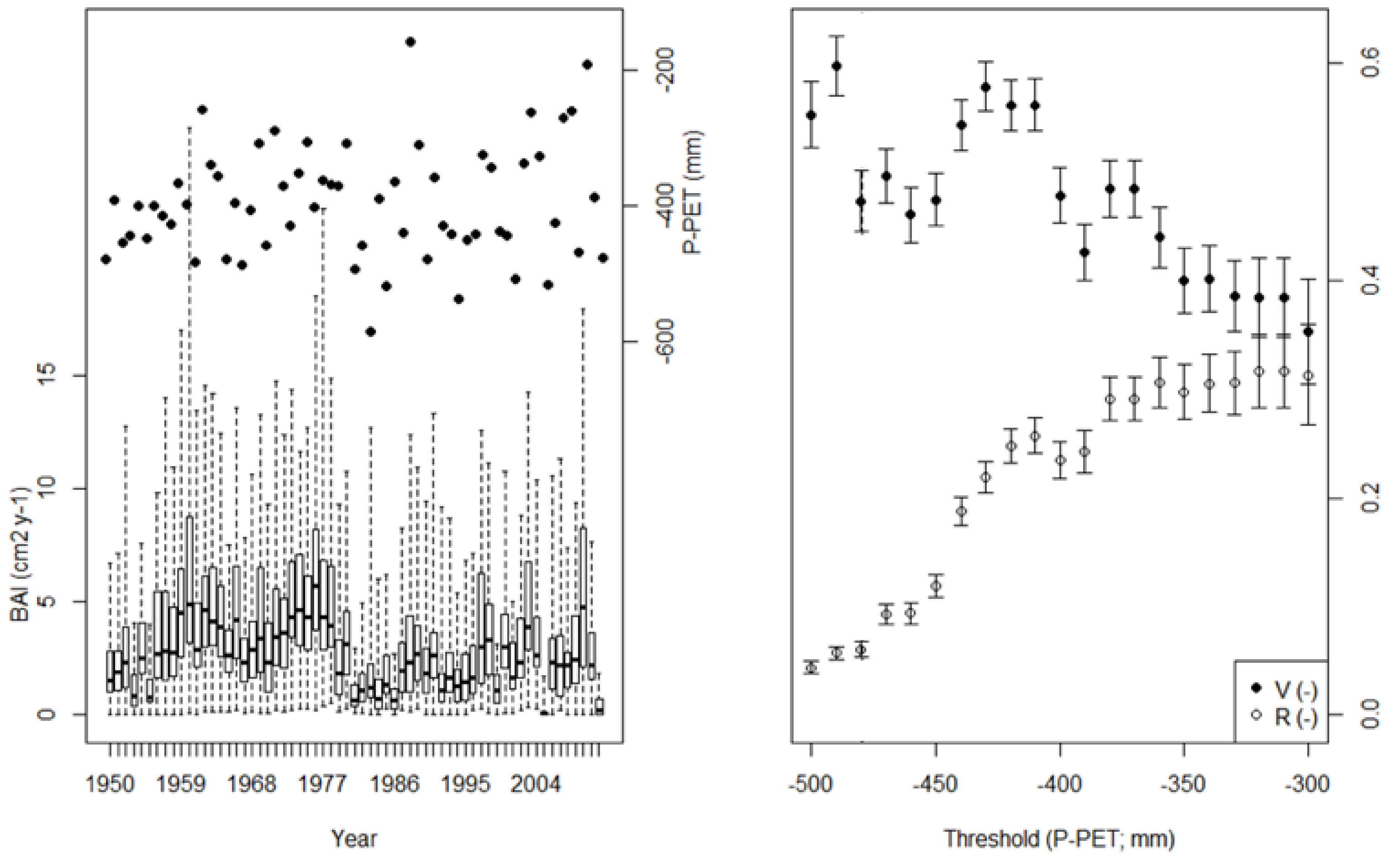
```
a. PRA(  x_r3,    -z_Pa,  250 )
b. PRA(  x_r3,     z_Pa,  250 )
c. PRA( -x_r3,     z_Pa, -250 )
```

If you want barplots, you can wrap the lines in `plotPRA( … )`

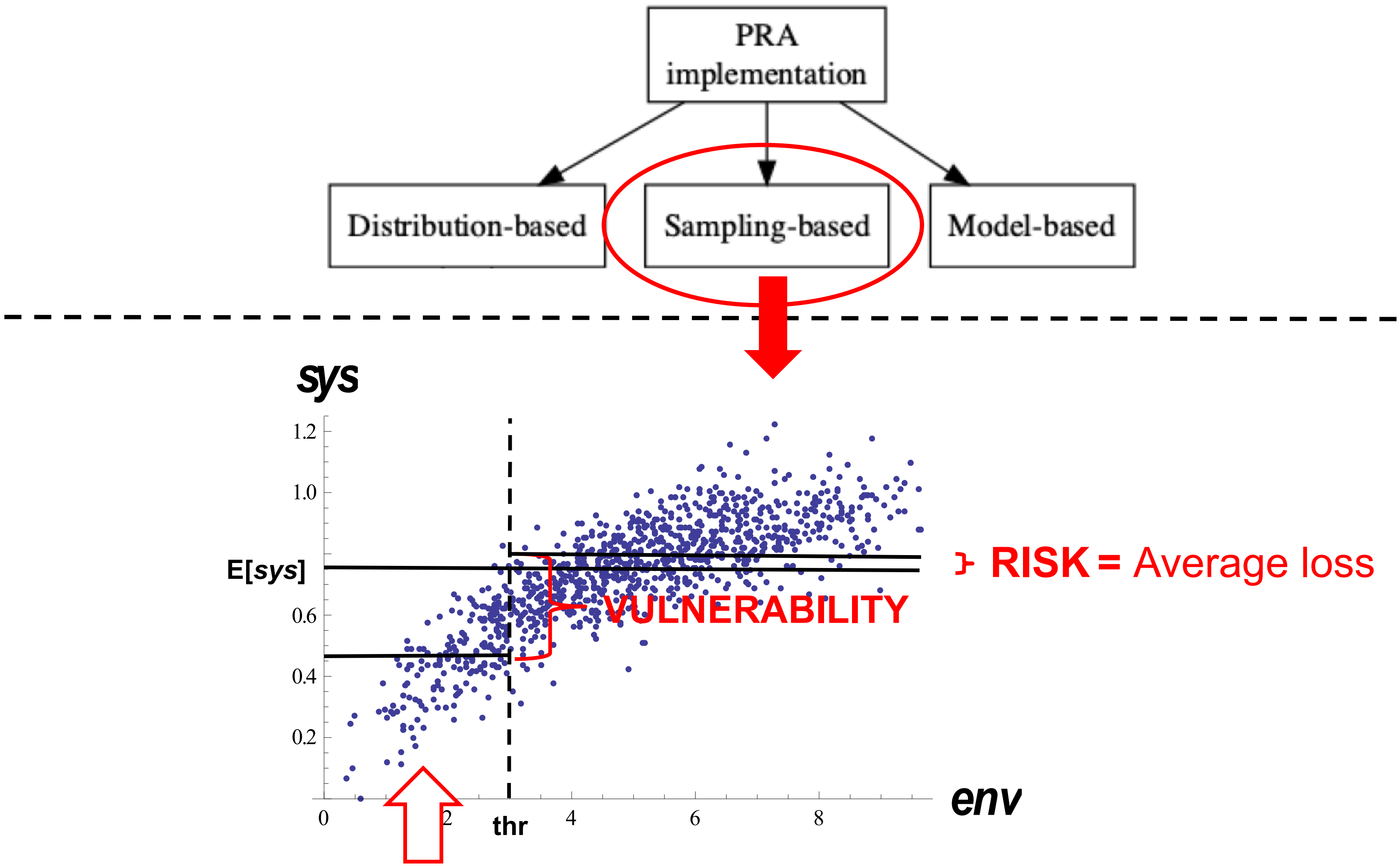
Forest data from Germany: EXERCISE 3

Change the R-code: try other choices for the threshold. How does that affect the results?

Van Oijen & Zavala (2019): P. sylvestris (El Carrascal, Spain)

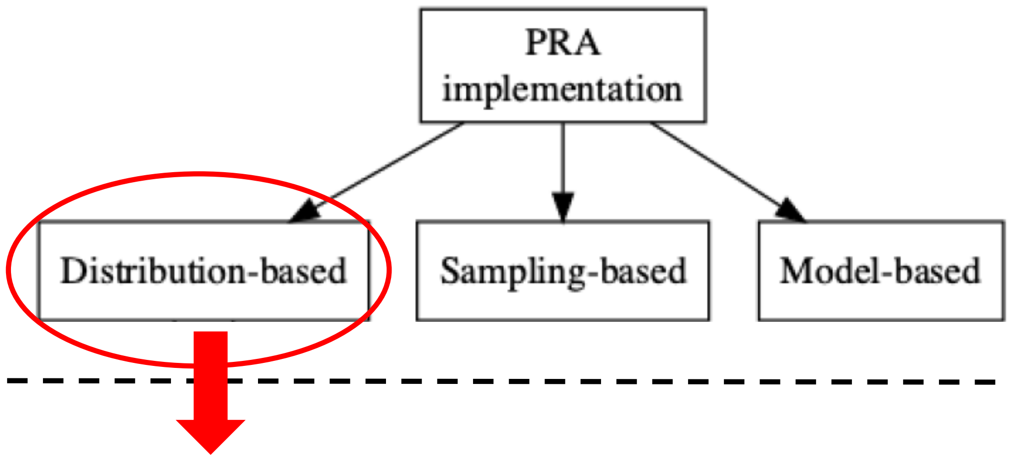


Deriving p[H], V and R from data (3 methods)



P(Hazardous) =
frequency *env* < thr =
n_H / n

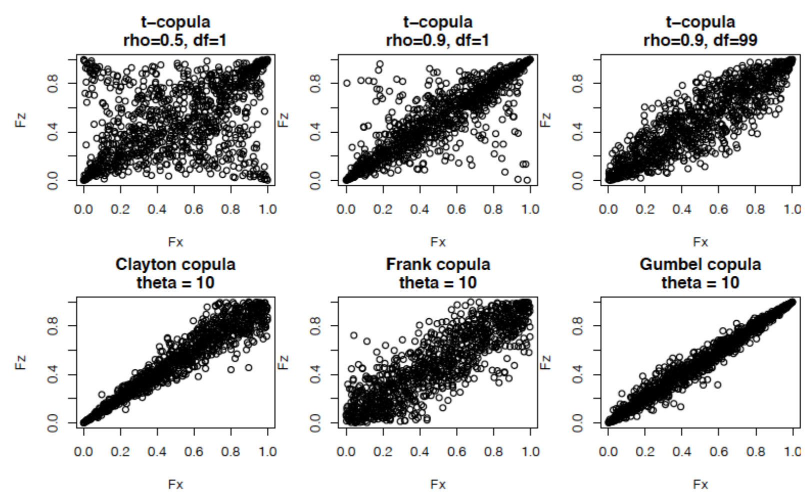
Deriving p[H], V and R from data (3 methods)



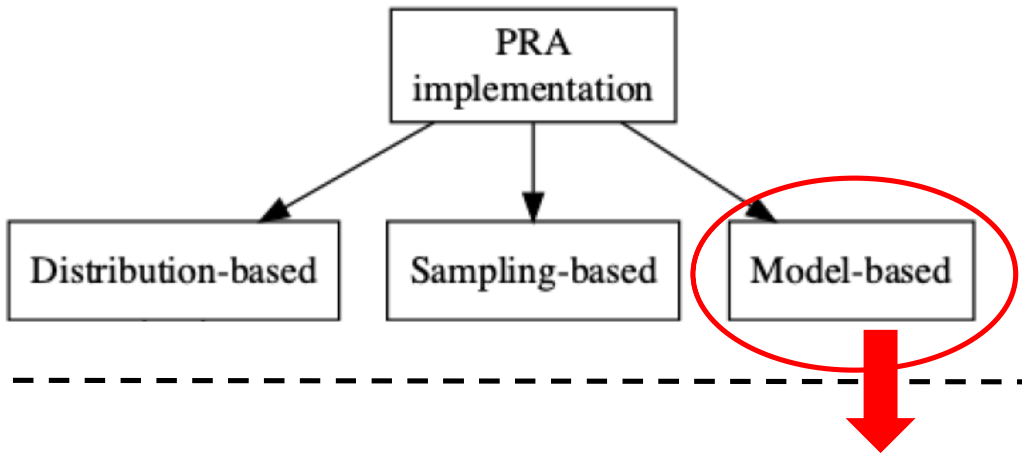
1. Fit a <u>distribution</u> p[x,z] to the data (*)
2. Integrate conditional expectations E[z|x]
3. Derive p[H], V and R

(*) *<u>Copulas</u>* can map any p[x] and p[z] to p[x,z]
⇒ freedom to choose appropriate probability distributions for x and z.

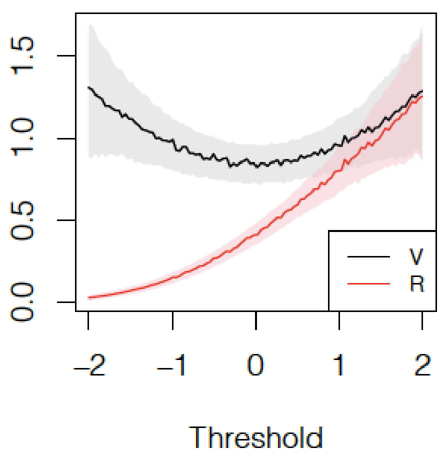
Semi-Bayesian copula-selection in R: function *BiCopSelect* from package *VineCopula*



Deriving p[H], V and R from data (3 methods)



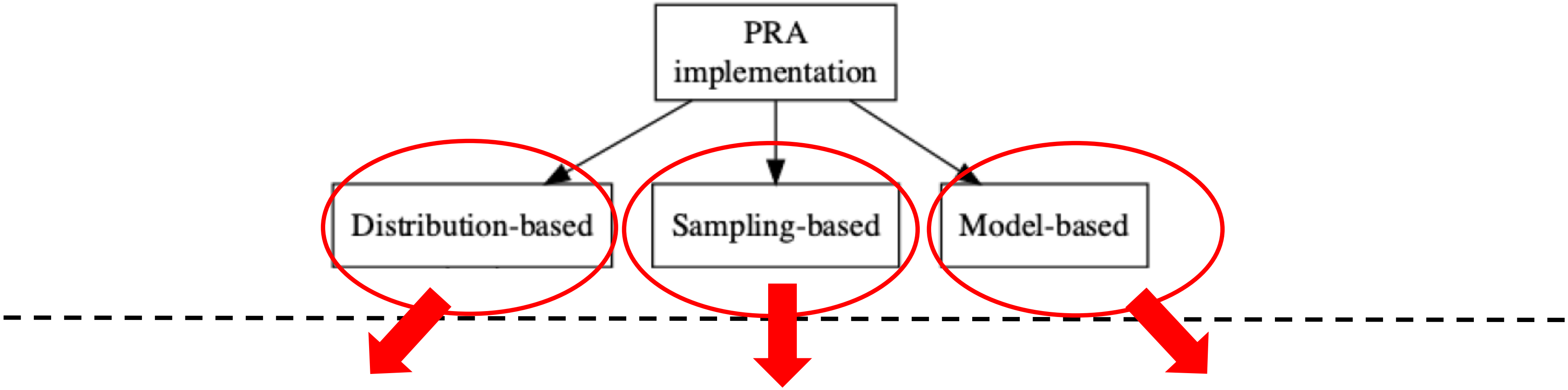
1. Fit a <u>model</u> $z = f(x,\theta)$ to the data (*)
2. Calculate conditional expectations $E[z|f(x,\theta)]$
3. Derive p[H], V and R



(*) <u>Bayesian</u> <u>calibration</u> estimates the model's parameters θ with uncertainty quantified

We often use R-package 'Nimble'

UQ



Distributional variances associated with p[x,z]

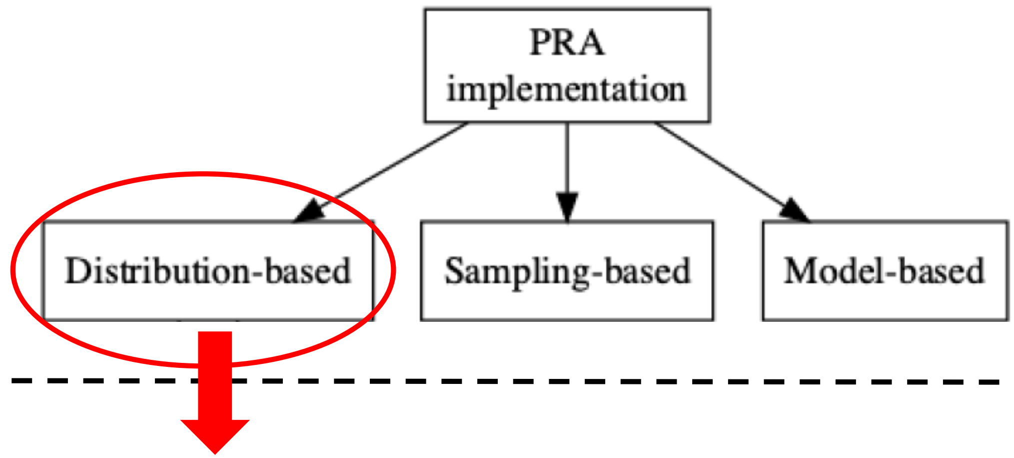
Sampling variances

$$\sigma_{\hat{E}[z|x<thr]} = \sqrt{\frac{Var[z_H]}{n_H}},$$

$$\sigma_{\hat{E}[z|x\geq thr]} = \sqrt{\frac{Var[z_{\neg H}]}{n-n_H}}.$$

Forward propagation of posterior parameter variances to model output variances

Distribution-based PRA



1. Fit a distribution p[x,z] to the data
2. Integrate conditional expectations E[z|x]
3. Derive p[H], V and R

Three cases:

1. Distribution known exactly
2. Distribution known, apart from one or more hyperparameters (e.g. bivariate Gaussian with unknown mean and covariance matrix)
3. Distribution unknown (e.g. major uncertainty about tails)

Conditional means

All probability distributions p[x,z] :

$$E[z|x < thr] = \frac{1}{F_x[thr]} \int_{x=-\infty}^{thr} p[x]\, E[z|x]\, dx$$

$$E[z|x \geq thr] = \frac{1}{1 - F_x[thr]} \int_{x=thr}^{\infty} p[x]\, E[z|x]\, dx$$

$$p[H] = F_x[thr]$$

Bivariate Gaussian p[x,z] = *N*[*m*, *S*] :

$$E[z|x < thr] = E[z] - \frac{Cov(x,z)\, p_x[thr]}{F_x[thr]}$$

$$E[z|x \geq thr] = E[z] + \frac{Cov(x,z)\, p_x[thr]}{1 - F_x[thr]}$$

Distribution-based PRA: Bivariate Gaussian

$$p[H] = F_x[thr]$$
$$V = \frac{Cov(x,z)\, p_x[thr]}{F_x[thr]\,(1 - F_x[thr])}$$
$$R = p[H]\, V$$

```
PRA0_Gauss <- function( m., S., thr. ) {
mx   <- m.[1] ; sx <- sqrt(S.[1,1]) ; Vxz <- S.[1,2]
pH   <- pnorm( thr., mx, sx )
V    <- Vxz * dnorm(thr., mx, sx) / (pH * (1-pH))
R    <- pH * V
return( c(pH=pH, V=V, R=R) ) }
```

But how to quantify uncertainty in distribution-based PRA?

We need to quantify uncertainty about hyperparameters

- Bivariate Gaussian: uncertainty about m and S

Bayes' Theorem

x=A: Infected
x=a: Not infected

p[x]

p[z|x]

Say we observe a value for z
What is **p[x|z]**?

$$p[x,z] = p[x]\, p[z|x]$$
$$= p[z]\, p[x|z]$$
$$\Rightarrow \quad \mathbf{p[x|z] = p[x]\, p[z|x] / p[z]}$$

"Bayes' Theorem"

Law of Total Probability (LTP)

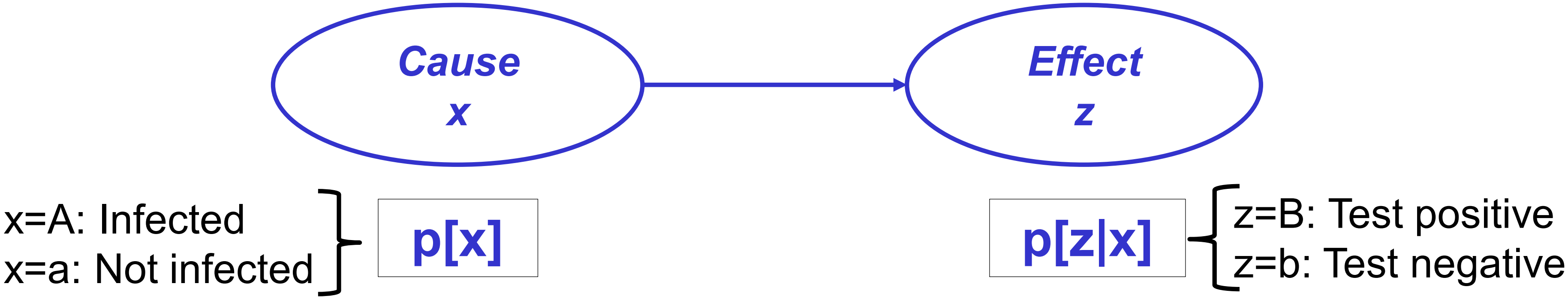
Binary x: $p[z] = p[x_0]\, p[z|x_0] + p[x_1]\, p[z|x_1]$

Discrete x: $p[z] = \sum p[x_i]\, p[z|xi]$

Continuous x: $p[z] = \int p[x]\, p[z|x]\, dx$

Often OK to ignore p[z] and write $\boldsymbol{p[x|z] \propto p[x]\, p[z|x]}$

Example - EXERCISE

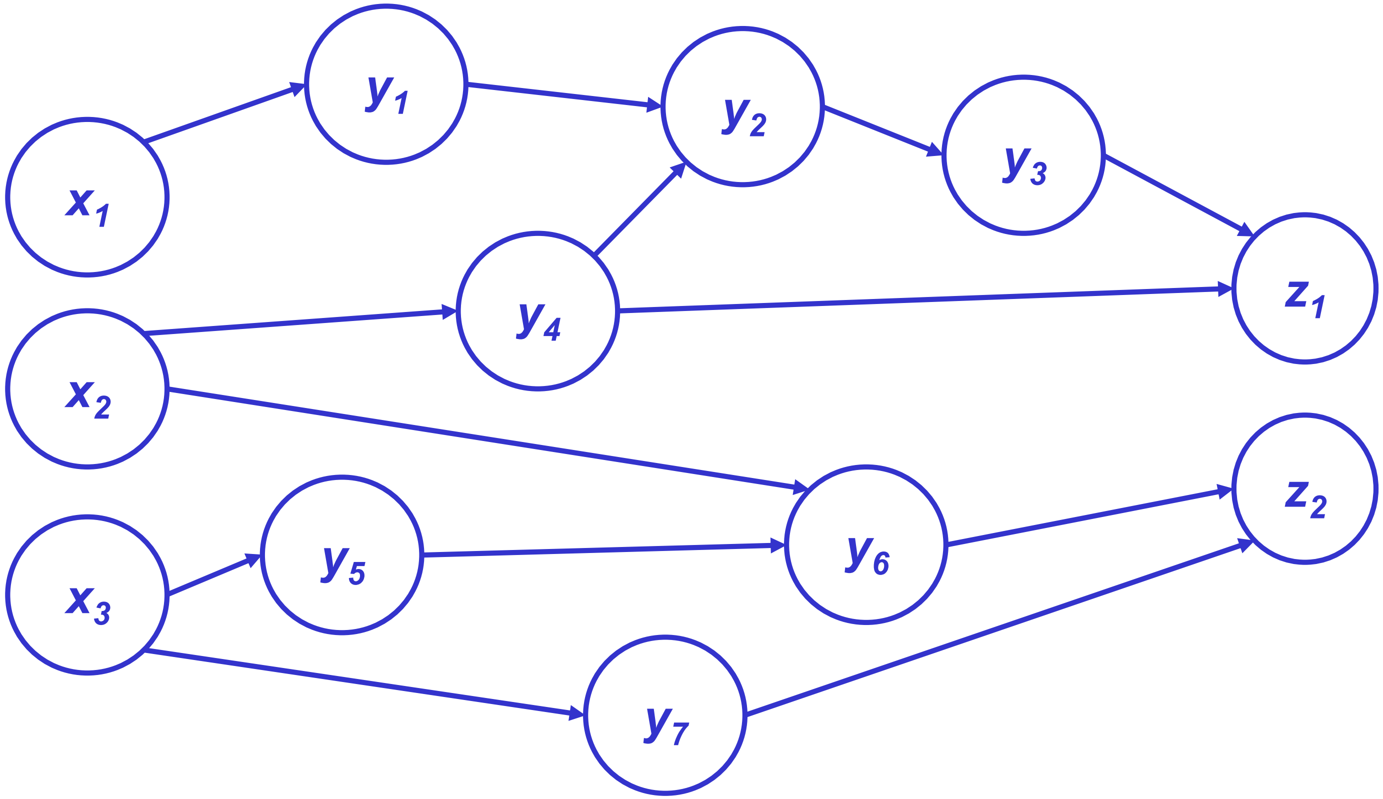


- Low prior infection probability: $p[A] = 0.01$
- Reliable test: $p[B|A] = p[b|a] = 0.99$

- Test result: B
- What is $p[A|B]$?

- Binary x ⇒ $p[B] = 0.01 \times 0.99 + 0.99 \times 0.01$
- Bayes: $p[A|B] = 0.01 \times 0.99 / p[B] = 0.5$

A complex joint probability distribution

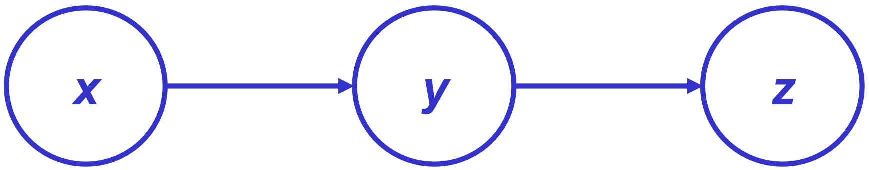


Updating the whole network:

- If given x_i ⇒ Forward propagation of uncertainty
- If given z_i ⇒ Bayes' Theorem
- If given y_i ⇒ *Both methods*

Prior, Law of Total Probability, Bayes' Theorem

$x \in \{A,a\}$ $y \in \{B,b\}$ $z \in \{C,c\}$



p[A]=0.9

p[B|A] = 0.9
p[B|a] = 0.6

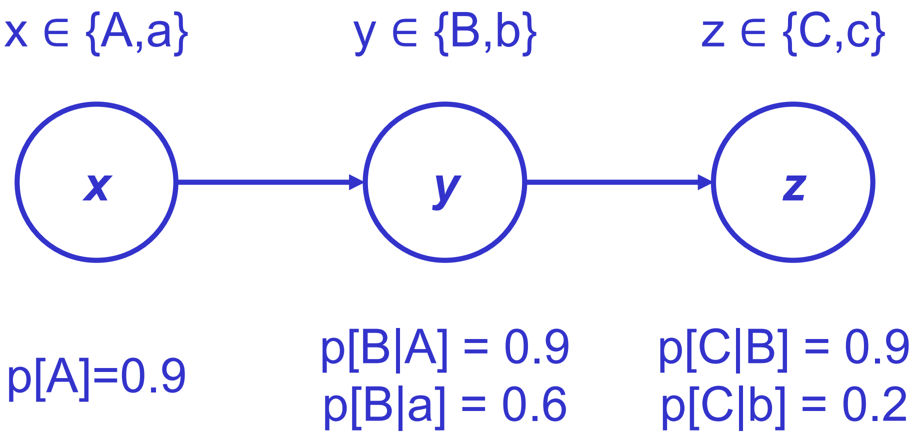
p[C|B] = 0.9
p[C|b] = 0.2

QUESTIONS

1. What are p[B] and p[C]?
2. Say x = a. What are p[B|a] and p[C|a]?
3. Say y = b. What are p[A|b] and p[C|b]?
4. Say z = c. What are p[A|c] and p[B|c]?

LTP	1. p[B] = p[A] p[B|A] + (1-p[A]) p[B|a] = 0.9×0.9 + 0.1×0.6 = 0.87
LTP	p[C] = p[B] p[C|B] + (1-p[B]) p[C|b] = 0.87×0.9 + 0.13×0.2 = 0.809
Prior	2. p[B|a] = 0.6
LTP	p[C|a] = p[B|a] p[C|B] + (1-p[B|a]) p[C|b] = 0.6×0.9 + 0.4×0.2 = 0.62
BT	3. p[A|b] = p[A] (1-p[B|A]) / (1-p[B]) = 0.9×0.1 / 0.13 = 0.692
Prior	p[C|b] = 0.2
BT	4. p[A|c] = p[A] p[c|A] / (1-p[C]) = 0.9*0.17 / 0.191 = 0.801
LTP	because p[c|A] = p[B|A] (1-p[C|B]) + (1-p[B|A]) (1-p[C|b]) = 0.17
BT	p[B|c] = p[B] (1-p[C|B]) / (1-p[C]) = 0.87×0.1 / 0.191 = 0.455

Prior, Law of Total Probability, Bayes' Theorem



QUESTIONS

1. What are p[B] and p[C]?
2. Say x = a. What are p[B|a] and p[C|a]?
3. Say y = b. What are p[A|b] and p[C|b]?
4. Say z = c. What are p[A|c] and p[B|c]?

```
pA   <- 0.9 ; pB_A <- 0.9 ; pB_a <- 0.6 ; pC_B <- 0.9 ; pC_b <- 0.2
n    <- 1e6
x    <- rbinom( n, 1, pA )
y    <- rbinom( n, 1, x*pB_A + (1-x)*pB_a )
z    <- rbinom( n, 1, y*pC_B + (1-y)*pC_b )

pB   <- sum(y==1) / n ; pC <- sum(z==1) / n
pB_a <- sum(x==0 & y==1) / sum(x==0)
pC_a <- sum(x==0 & z==1) / sum(x==0)
pA_b <- sum(x==1 & y==0) / sum(y==0)
pC_b <- sum(y==0 & z==1) / sum(y==0)
pA_c <- sum(x==1 & z==0) / sum(z==0)
pB_c <- sum(y==1 & z==0) / sum(z==0)
```

```
# Analytical:
[1,] 0.870 0.809
[2,] 0.600 0.620
[3,] 0.692 0.200
[4,] 0.801 0.455
```

```
# Numerical:
[1,] 0.870 0.809
[2,] 0.598 0.617
[3,] 0.691 0.200
[4,] 0.800 0.457
```

Conjugate Bayesian updating

p[ϑ|y] = p[ϑ] p[y|ϑ] / p[y]

If p[ϑ] and p[y|ϑ] are conjugate, then p[ϑ|y] can be calculated analytically

Gaussian - Gaussian:

$$p[\theta] = N[\mu_0, \sigma_0^2] \; ; \; p[y|\theta] = N[\theta, \sigma_y^2] \quad \Longrightarrow \quad p[\theta|y] = N[\mu_1, \sigma_1^2]$$

$$\text{where} \quad \mu_1 = \frac{\mu_0\,\sigma_y^2 + y\,\sigma_0^2}{\sigma_y^2 + \sigma_0^2} \quad \text{and} \quad 1/\sigma_1^2 = 1/\sigma_0^2 + 1/\sigma_y^2$$

Beta - Binomial:

$$p[H] \sim Be[a,b] \; ; \; p[n_H|p[H]] = Bi[n, p[H]] \quad \Longrightarrow$$

$$p[H|n, n_H] = Be[a + n_h, b + n - n_H]$$

Dirichlet - Multinomial:

$$\mathbf{p}[\mathbf{H}] \sim Di_k[\alpha] \; ; \; p[\mathbf{n_H}|\mathbf{p}[\mathbf{H}]] = Mu_k[\mathbf{p}[\mathbf{H}]] \quad \Longrightarrow \quad \mathbf{p}[\mathbf{H}|\mathbf{n_H}] = Di_k[\alpha + \mathbf{n_H}]$$

$$\text{where} \quad \mathbf{p}[\mathbf{H}], \alpha, \mathbf{n_H} \quad \text{are vectors of length} \quad k$$

Conjugate Bayesian updating of bivariate Gaussian

Prior:

$$p[\mu, \Sigma] \propto |\Sigma|^{-3/2}$$

Data:

$$\mathbf{y} = \{\, \mathbf{n},\ \mathbf{m} = (\bar{x}, \bar{z}),\ \mathbf{S} = cov(x, z)\, \}$$

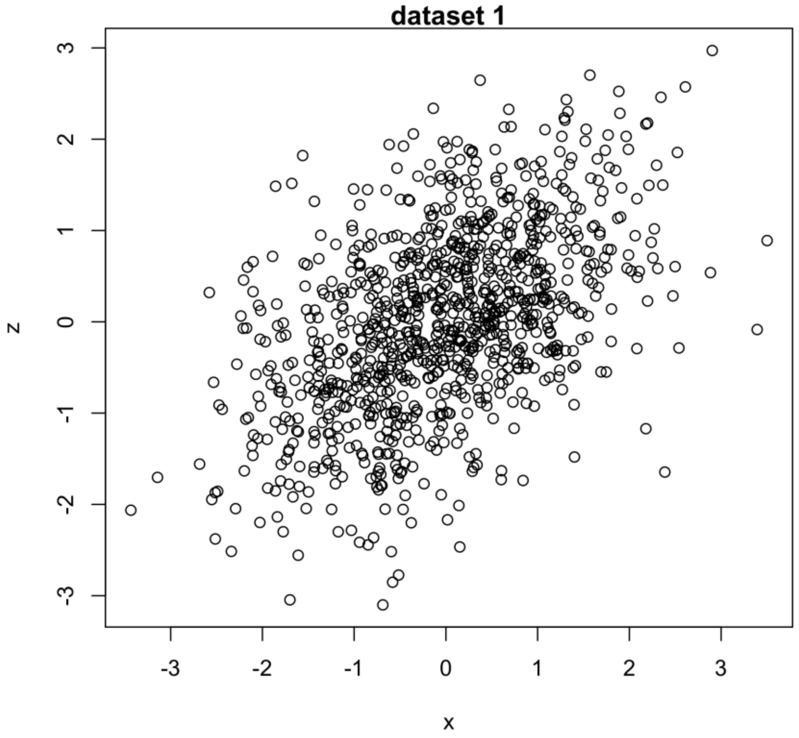
Posterior:

$$p[\Sigma|\mathbf{y}] = IWi_{\mathbf{n}-1}(\mathbf{S}'^{-1}), \text{ where } S' = (n-1)S$$

$$p[\mu|\Sigma, \mathbf{y}] = N[\mathbf{m}, \Sigma/\mathbf{n}].$$

```
PRA_Gauss <- function(m., S., n., thr.) {
  pH <- V <- R <- rep( NA, 1e3 )
  for(j in 1:1e3){
    S      <- riwish( n.-1, S. * (n.-1) )
    m      <- rmvnorm( 1, m., S/n. )
    PRA    <- PRA0_Gauss( m, S, thr. )
    pH[j] <- PRA["pH"] ; V[j] <- PRA["V"] ; R[j] <- PRA["R"]
  }
  return( c(   pH=mean(pH),   V=mean(V),   R=mean(R),
               s_pH=sd  (pH), s_V=sd  (V), s_R=sd  (R) ) )
```

Posterior inverse-Wishart distribution



$\Longrightarrow$

y = {n,m,S}

n = 1000
m = (-0.040,0.054)
**S = (1.129, 0.537
0.537, 1.021)**

$\Downarrow$

```
riwish( n-1, S*(n-1) )
```

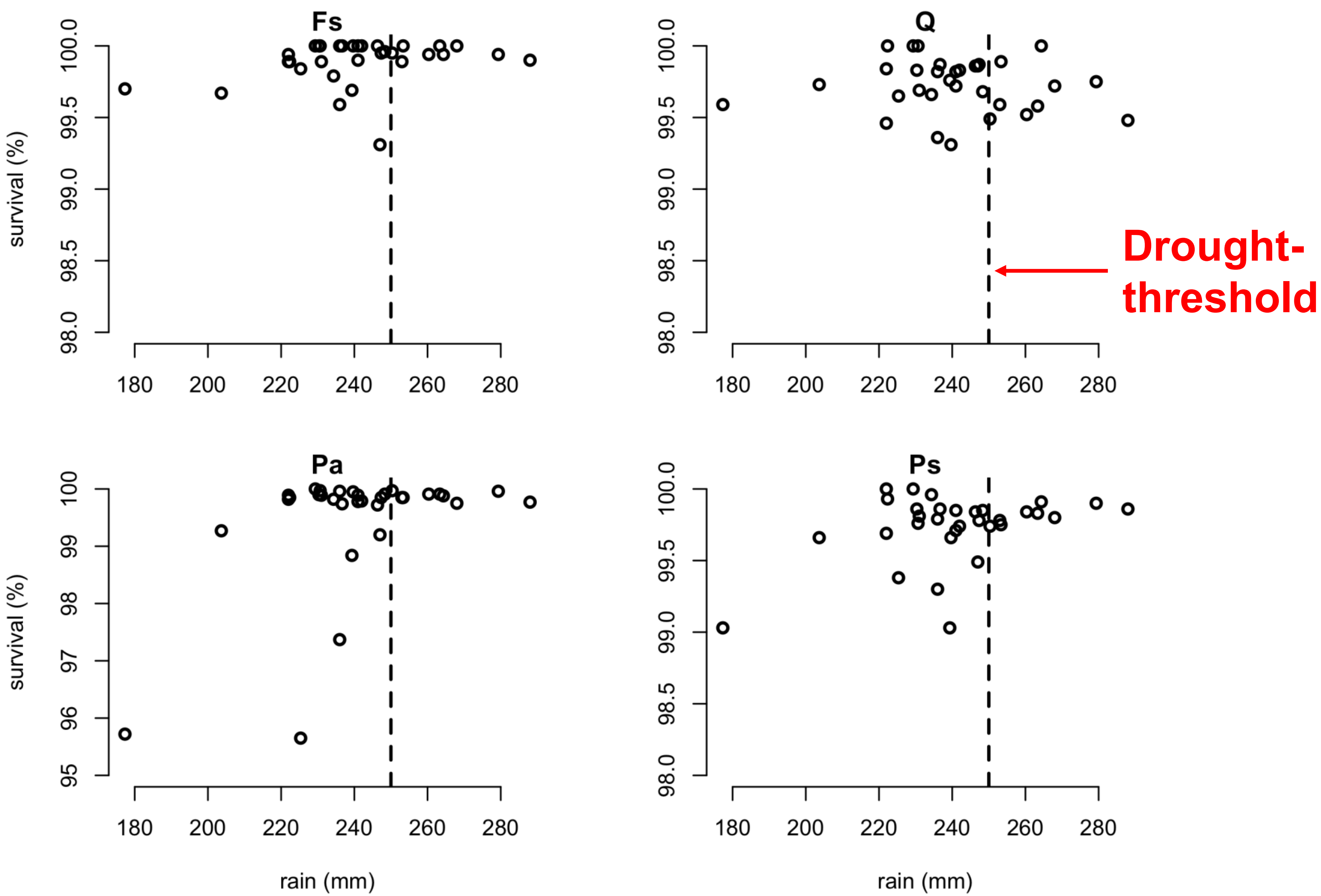
$\Downarrow$

$$\begin{bmatrix} 1.17 & 0.57 \\ 0.57 & 1.04 \end{bmatrix}, \begin{bmatrix} 1.09 & 0.53 \\ 0.53 & 1 \end{bmatrix}, \begin{bmatrix} 1.18 & 0.54 \\ 0.54 & 0.93 \end{bmatrix}, \ldots$$

Non-Bayesian UQ for Bivariate-Gaussian PRA

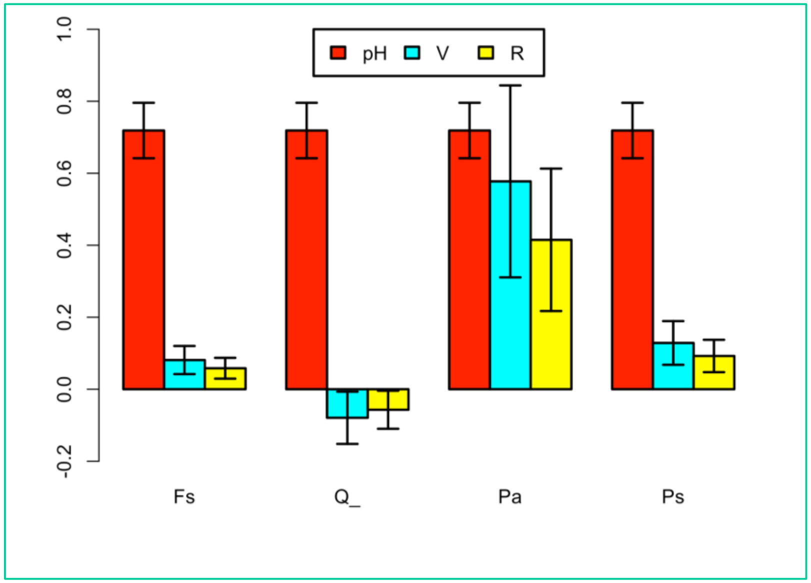
```
PRA_Gauss5 <- function( m., S., n., thr. ) {

  PRA  <- PRA0_Gauss( m., S., thr. )
  pH   <- PRA["pH"] ; V <- PRA["V"] ; R <- PRA["R"]

  s_pH <- sqrt( pH*(1-pH) / n. )
    r          <-  S[1,2] / sqrt( prod(diag(S)) )
    VarEz_H    <- (S[2,2] + r*(thr.-m[1])*(R-V) - (R-V)^2) / (n.*pH)
    VarEz_NotH <- (S[2,2] + r*(thr.-m[1])* R - R^2 ) / (n.*(1-pH))
  s_V  <- sqrt( VarEz_NotH + VarEz_H )
  s_R  <- sqrt( s_pH^2*s_V^2 + s_pH^2*V^2 + pH^2*s_V^2 )

  return( c( pH, V, R, s_pH, s_V, s_R ) )
}
```

Forest survival data from Germany

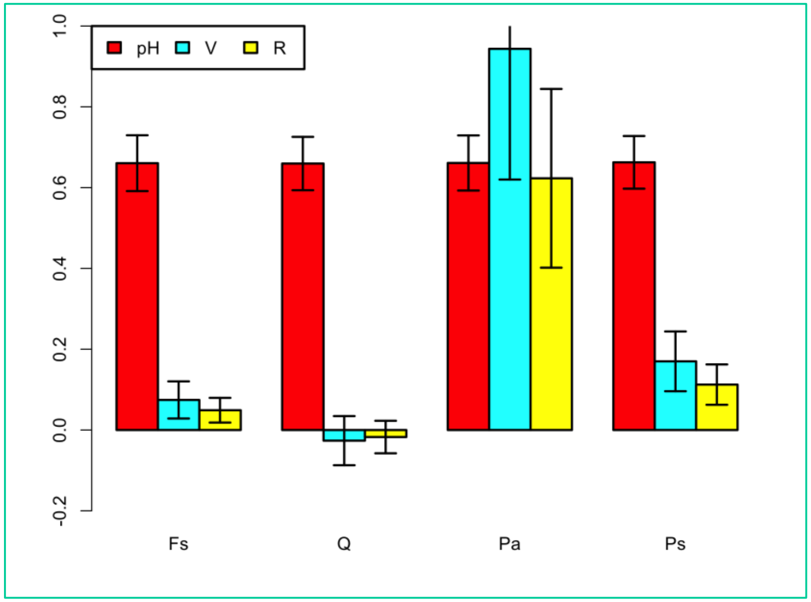


Sampling-based vs. distribution-based PRA

Sampling-based PRA



Distribution-based PRA



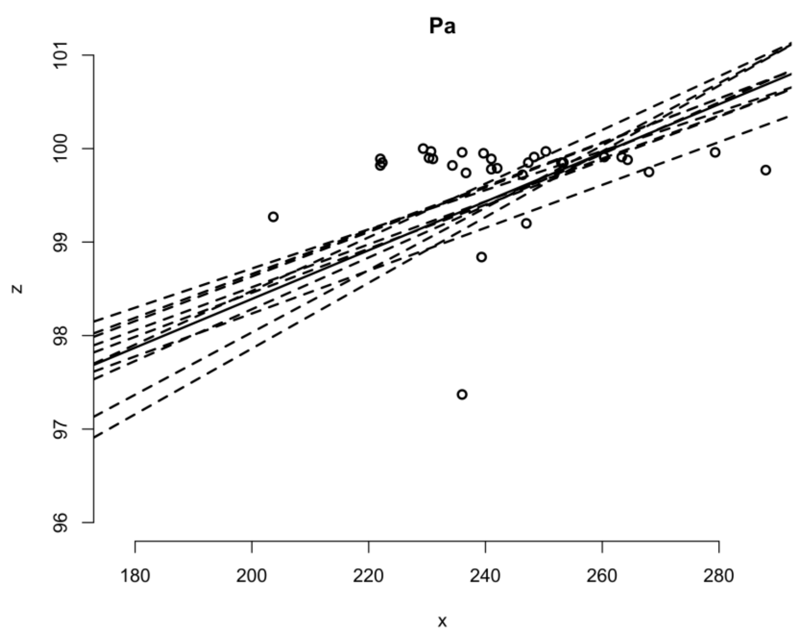
Conjugate Linear Bayes: Lindley & Smith (1972)

$$\text{Model: } y = \mathbf{X}\beta + \epsilon \quad \text{with } \beta \sim N[\mu_\beta, \Sigma_\beta], \epsilon \sim N[0, \Sigma_y]$$

$$\text{Posterior covariance matrix: } \Sigma_{\beta|y} = (\Sigma_\beta^{-1} + \mathbf{X}^\top \Sigma_y^{-1} \mathbf{X})^{-1}$$

$$\text{Posterior mean: } \mu_{\beta|y} = \Sigma_{\beta|y}(\Sigma_\beta^{-1}\mu_\beta + \mathbf{X}^\top \Sigma_y^{-1} y)$$

```
X          <- cbind( 1, x )
mb         <- c(0   ,0   ) ; nb <- length(mb)
Vb         <- c(1.e4,1.e4) ; Sb <- diag(Vb)
Sb_y_LS72 <- solve( solve(Sb) + t(X) %*% solve(Sy) %*% X ) ;
mb_y_LS72 <- Sb_y_LS72 %*% (solve(Sb) %*% mb + t(X) %*% solve(Sy) %*% y)
```

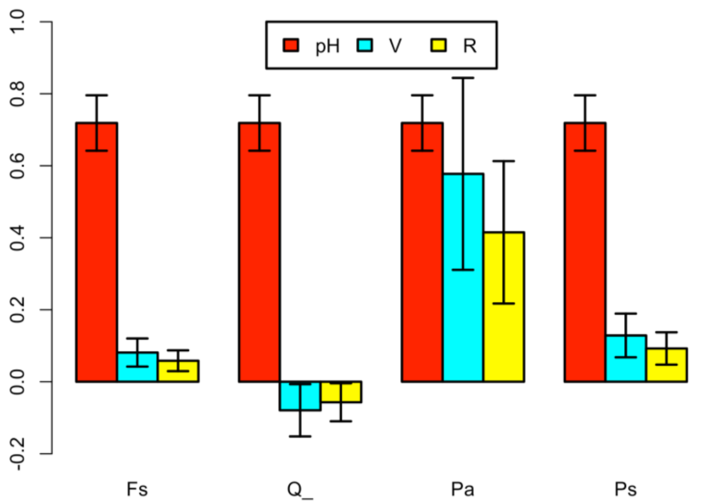


Linear-model-based PRA: code

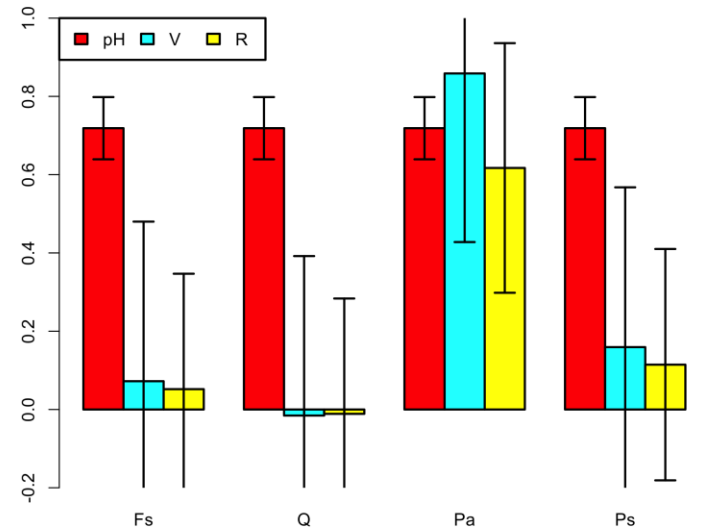
```
PRA_LS72 <- function( x, z, thr=0, Vz=1 ) {

  n          <- length(x)               ; X  <- cbind(1,x)
  mb         <- c(0,0)                  ; Vb <- c(1.e4,1.e4)
  Sb         <- diag(Vb)                ; Sz <- diag(Vz,n)
  Sb_y_LS72 <- solve( solve(Sb) + t(X) %*% solve(Sz) %*% X )
  mb_y_LS72 <- Sb_y_LS72 %*% (solve(Sb) %*% mb + t(X) %*% solve(Sz) %*% z)

  i_H      <- which( x <  thr )         ; n_H      <- length(i_H)
  i_NotH   <- which( x >= thr )         ; n_NotH   <- length(i_NotH)
  pH       <- n_H / n                   ; s_pH     <- sqrt( pH*(1-pH) / n )
  Ex_H     <- mean( x[i_H] )            ; Ex_NotH <- mean( x[i_NotH] )
  Ez_H     <- c(1,Ex_H) %*% mb_y_LS72 ; Ez_NotH <- c(1,Ex_NotH) %*% mb_y_LS72
  V        <- Ez_NotH - Ez_H            ; R        <- pH * V
    Vzi    <- function(i){ t(c(1,x[i])) %*% Sb_y_LS72 %*% c(1,x[i]) + Vz }
* Vz_H     <- sum( sapply(i_H   ,Vzi) ) / n_H    + mb_y_LS72[2]^2 * var(x[i_H])
* Vz_NotH <- sum( sapply(i_NotH,Vzi) ) / n_NotH + mb_y_LS72[2]^2 * var(x[i_NotH])
  s_V      <- sqrt( Vz_H / n_H + Vz_NotH / n_NotH )
  s_R      <- sqrt( s_pH^2 * s_V^2 + s_pH^2 * V^2 + pH^2 * s_V^2 )

  return( list( mb  = mb_y_LS72, Sb = Sb_y_LS72,
                PRA = c( pH=pH, V=V, R=R, s_pH=s_pH, s_V=s_V, s_R=s_R ) ) )
}
```

* By the law of total variance: $Var[z] = E[Var[z|x]] + Var[E[z|x]]$

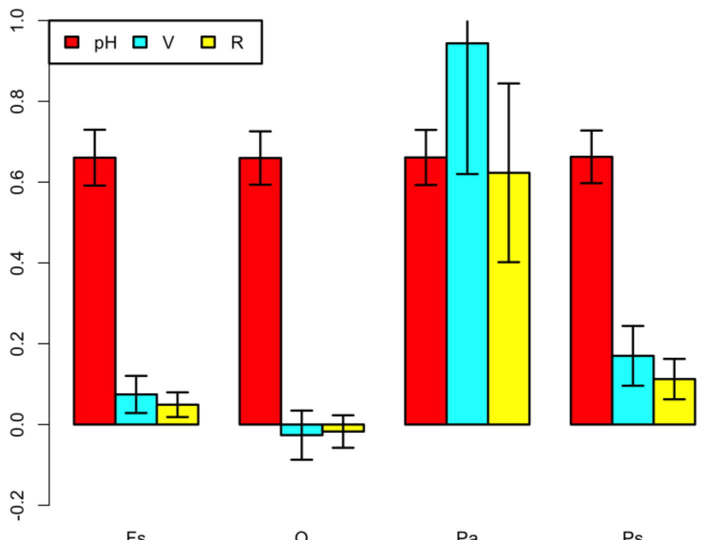
Sampling-, Distribution-, and Model-based PRA



Sampling-based PRA



Model-based PRA

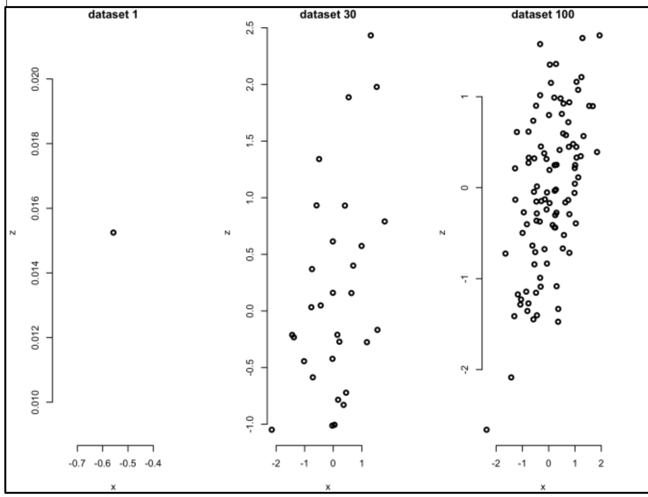


Distribution-based PRA

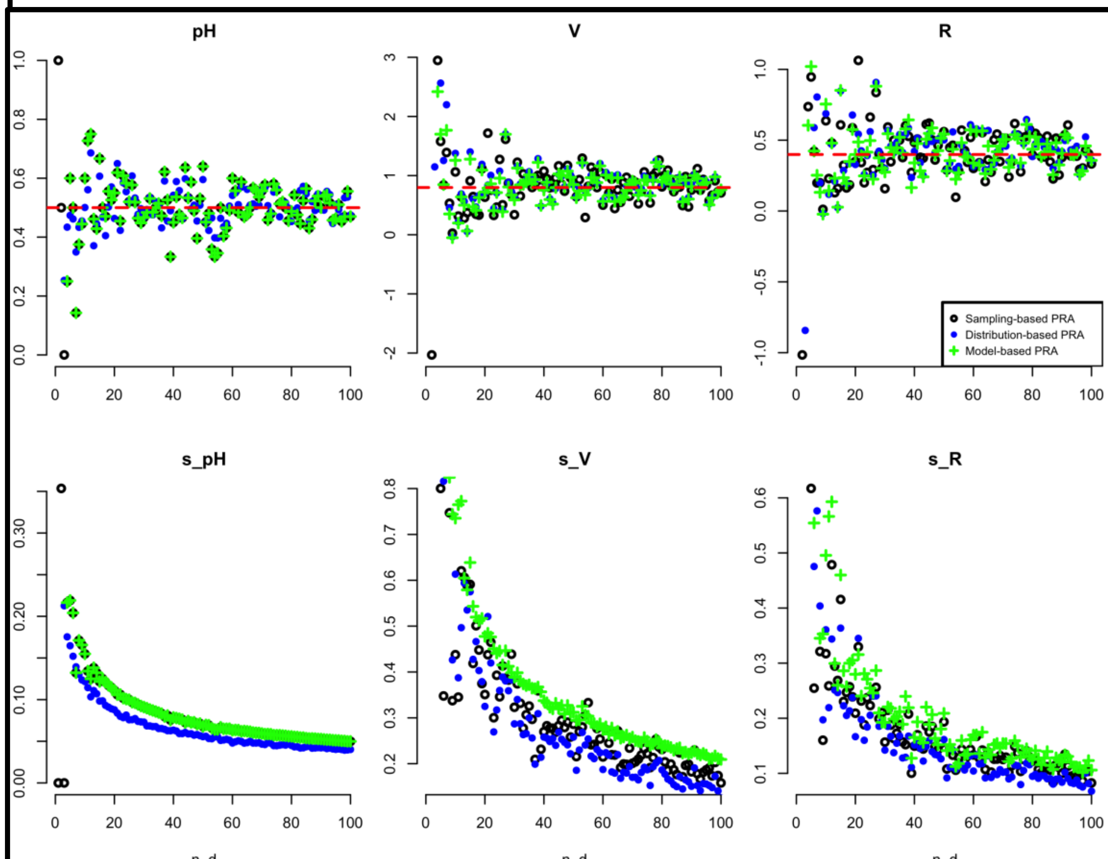
Which PRA-method would you choose?

PRAs on 100 ‘linear’ datasets (n = 1 : 100)

```
mu  <- c(0,0) ; Sigma <- diag(1,2) ; Sigma[1,2] <- Sigma[2,1] <- 0.5
n_d <- 1e2    ; l_xz.L2 <- vector("list",n_d)
for(d in 1:n_d) { l_xz.L2[[d]] <- rmvnorm( d, mu, Sigma ) }
```

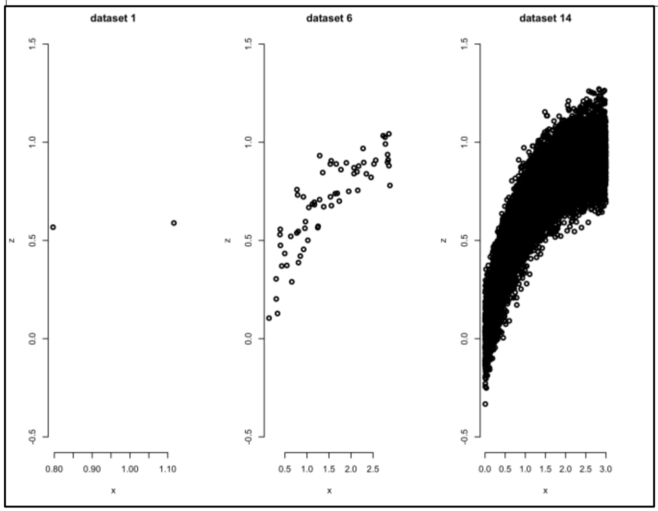



```
sapply( 1:n_d, function(d){
  PRA( l_xz.L2[[d]][,1], l_xz.L2[[d]][,2], thr=0 ) } )
```

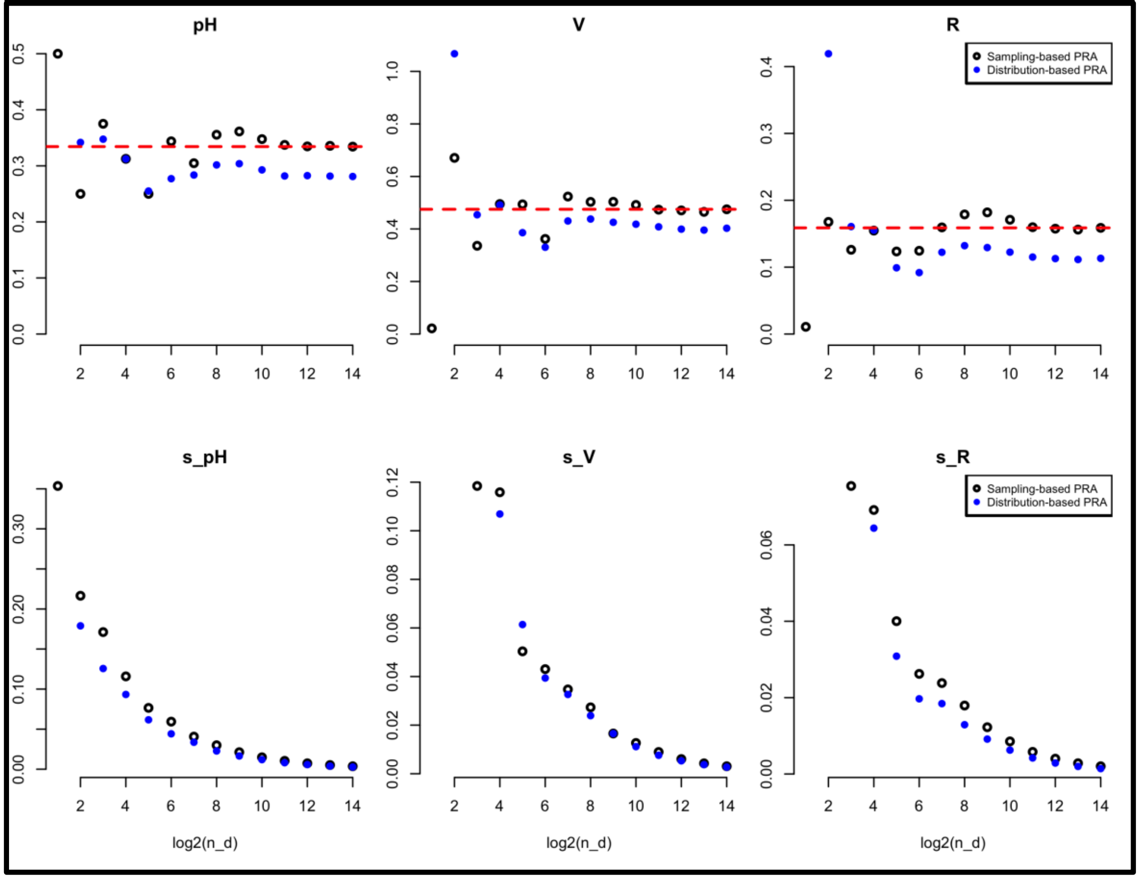


PRAs on 14 'nonlinear' datasets (n = 2 : 2^{14})

```
n_d <- 14 ; l_xz.NL3 <- vector("list",n_d) ; sz <- 0.1
for(d in 1:n_d){ x  <- runif(2^d,0,3)
                 ez <- rnorm(2^d,0,sz) ; z <- 1-exp(-x) + ez
                 l_xz.NL3[[d]] <- cbind(x,z) }
```




```
sapply( 1:n_d, function(d){
  PRA( l_xz.NL3[[d]][,1], l_xz.NL3[[d]][,2], thr=1 ) } )
```



USGS CC PDM 1.0

2.
Model-based PRA