Gwyddor data ar frig y don

Philip Jonathan

Prifysgol Caerhirfryn, Adran Fathemateg & Ystadegaeth
Shell Research Ltd., Llundain

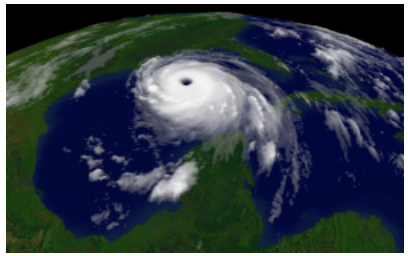
Cymdeithas Wyddonol Caerdydd, Tachwedd 2021

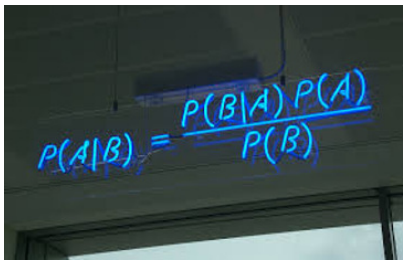
Braslun

- Cefndir personol
- Paham fod modelu ystadegol yn gymaint o **hwyl**?
- Gair am wyddor data
- **Eithafon morol**

Rhuthun

Godre'r Graig ac Abertawe

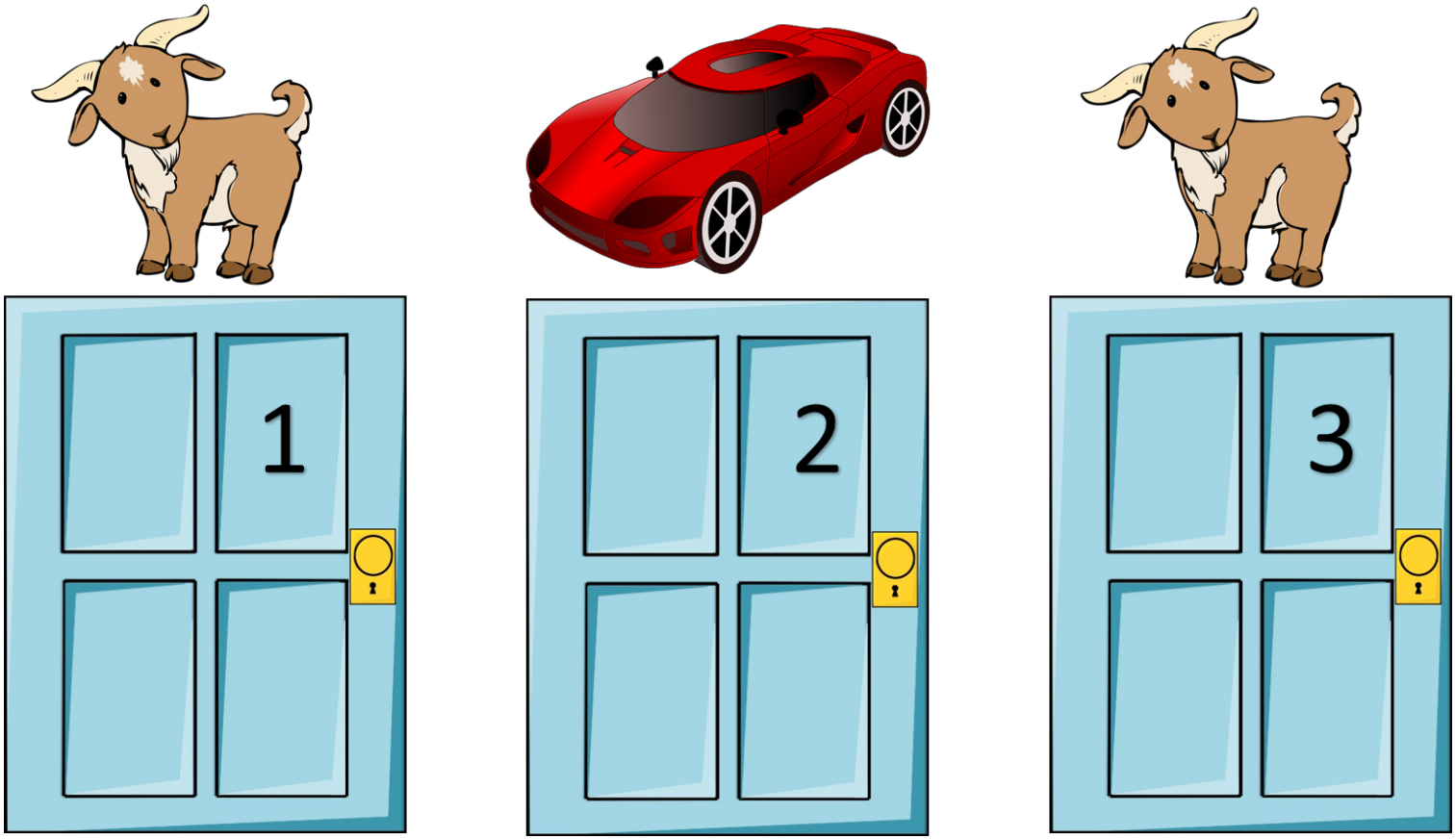




Corwynt Katrina (2005) a theorem Bayes

Shell Centre a Chaerhirfryn

Gafr neu Mazerati?



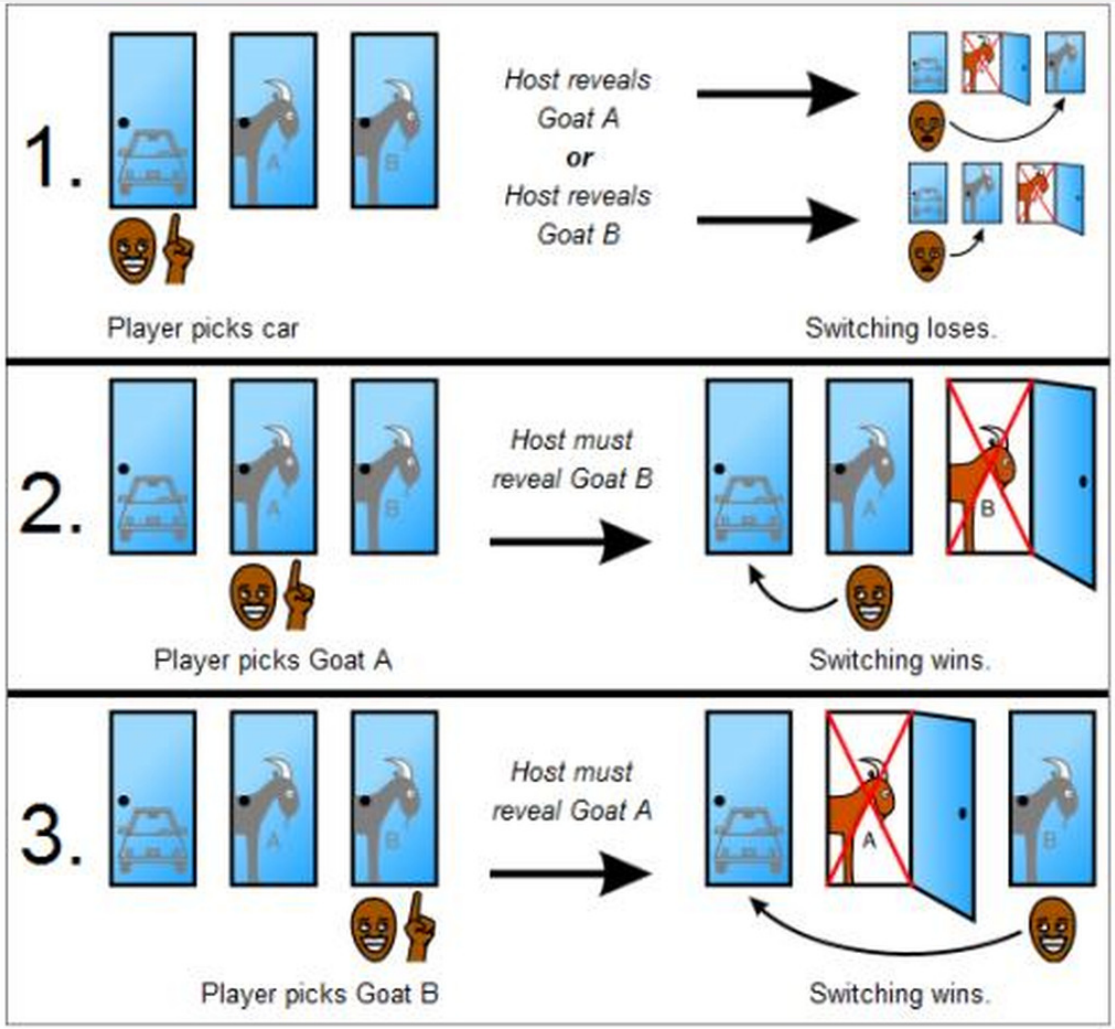
Gafr neu Mazerati?

Dychmygwch ...

- ’Da chi mewn cwis ar y teledu
- Ma’ ’da chi ddewis o dri drws
- Tu ôl i un drws mae ’na gar ffansi (rhywbeth dymunol iawn!); tu ôl i’r ddau drws arall mae ’na afr (neu rhywbeth ych-a-fi!).
- ’Da chi’n dewis drws, ac yn rhoi rhif y drws i’r boi teledu
- Mae’r boi teledu yn agor un o’r ddau drws na ddewisoch sydd â gafr y tu ôl iddo
- Mae’r boi teledu’n gofyn **Ydych chi eisiau newid eich dewis?**

- Be’ ’da chi’n gwneud? Ydi e o fantais i chi, i newid eich dewis?

Gafr neu Mazerati?

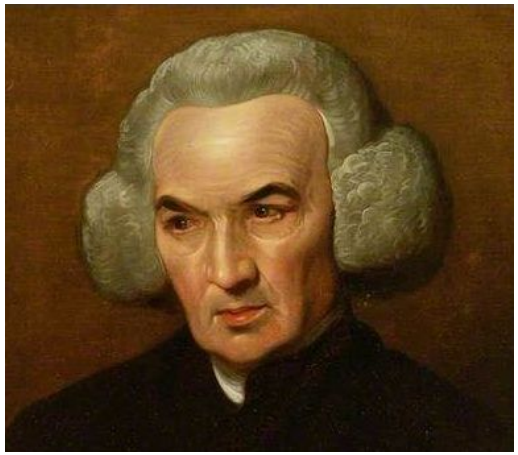


... weithiau, doethach newid meddwl

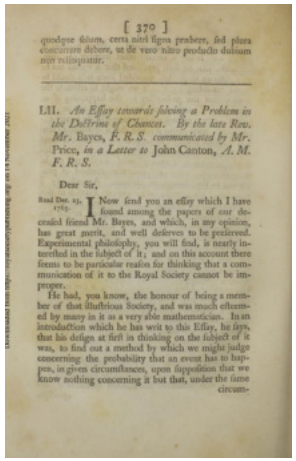
Tebygolrwydd (amodol) a theorem Bayes

$$p(B|D) = \frac{p(D|B) \times p(B)}{p(D)}$$

cred ddiweddarach $\propto$ tebygoliaeth y data $\times$ cred gychwynol



Richard Price (Llangeinwyr, 1723 – 1791)



[370]

quodque ſolum, certa nitri ſigna prabere, ſed plura concurrere debere, ut de vero nitro productio dubium non relinquatur.

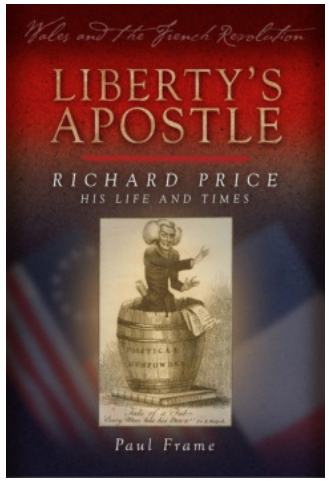
LII. *An Eſſay towards ſolving a Problem in the Doctrine of Chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a Letter to* John Canton, *A. M. F. R. S.*

Dear Sir,

Read Dec. 23, 1763. I Now ſend you an eſſay which I have found among the papers of our deceaſed friend Mr. Bayes, and which, in my opinion, has great merit, and well deſerves to be preſerved. Experimental philoſophy, you will find, is nearly intereſted in the ſubject of it; and on this account there ſeems to be particular reaſon for thinking that a communication of it to the Royal Society cannot be improper.

He had, you know, the honour of being a member of that illuſtrious Society, and was much eſteemed by many in it as a very able mathematician. In an introduction which he has writ to this Eſſay, he ſays, that his deſign at firſt in thinking on the ſubject of it was, to find out a method by which we might judge concerning the probability that an event has to happen, in given circumſtances, upon ſuppoſition that we know nothing concerning it but that, under the ſame circum-

Ei erthygl ar waith Thomas Bayes



Llyfr diweddar amdano

- Rhoi'r debygoliaeth a'r rhag-ddwysedd "fel rhifau" a'u "lluosi"!
- Sail ar gyfer dysgu ystadegol

Gwyddor data

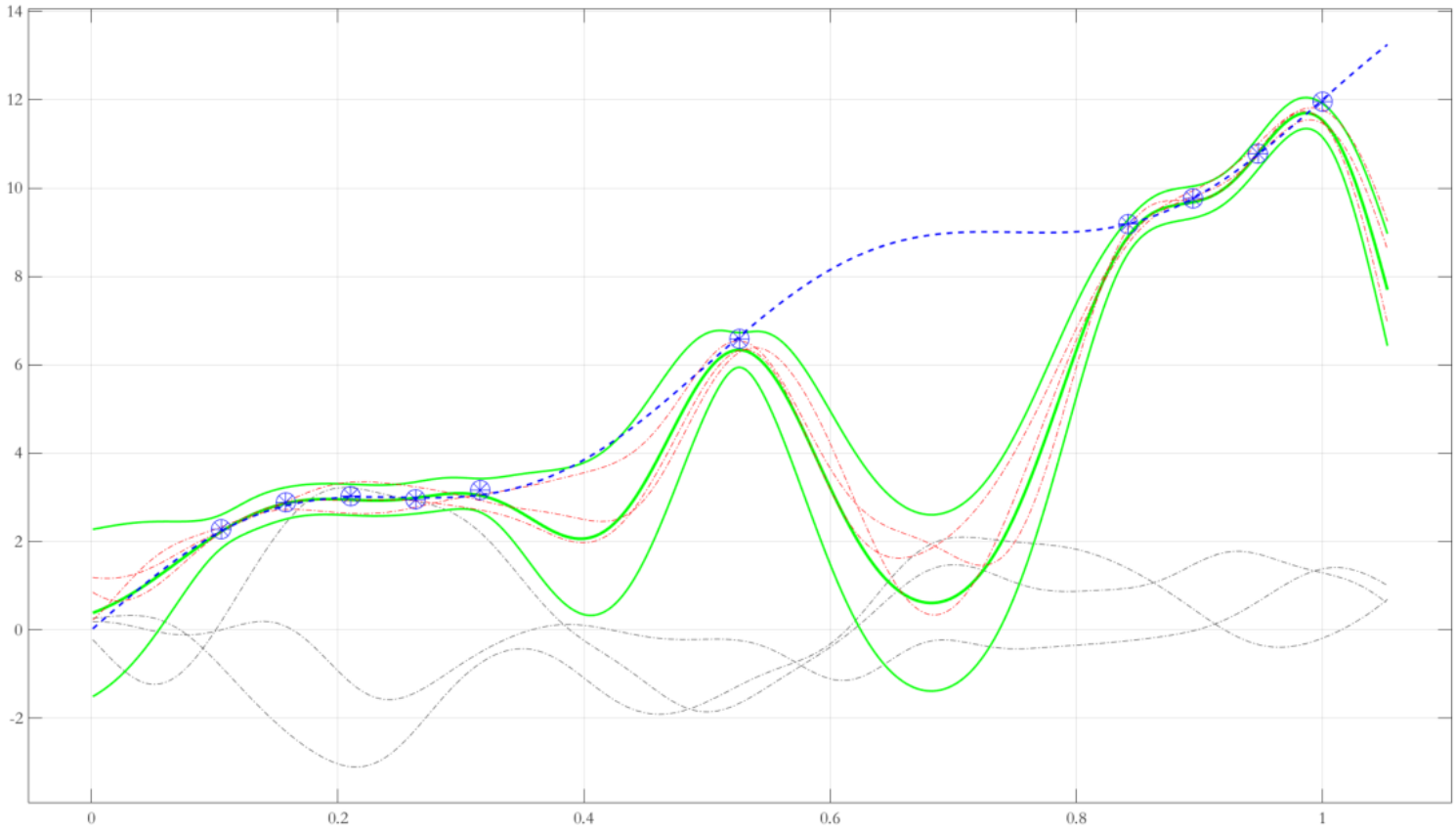
Panasea newydd?

- Yn ôl wiki, 4ydd paradeim gwyddoniaeth
- Empirig
- Damcaniaethol
- Cyfrifianol
- "Sail-data" (*data-driven*)

... ond mae'r sylfeini yr un peth ag erioed!

Gwyddor data, rhyngosod ac allosod

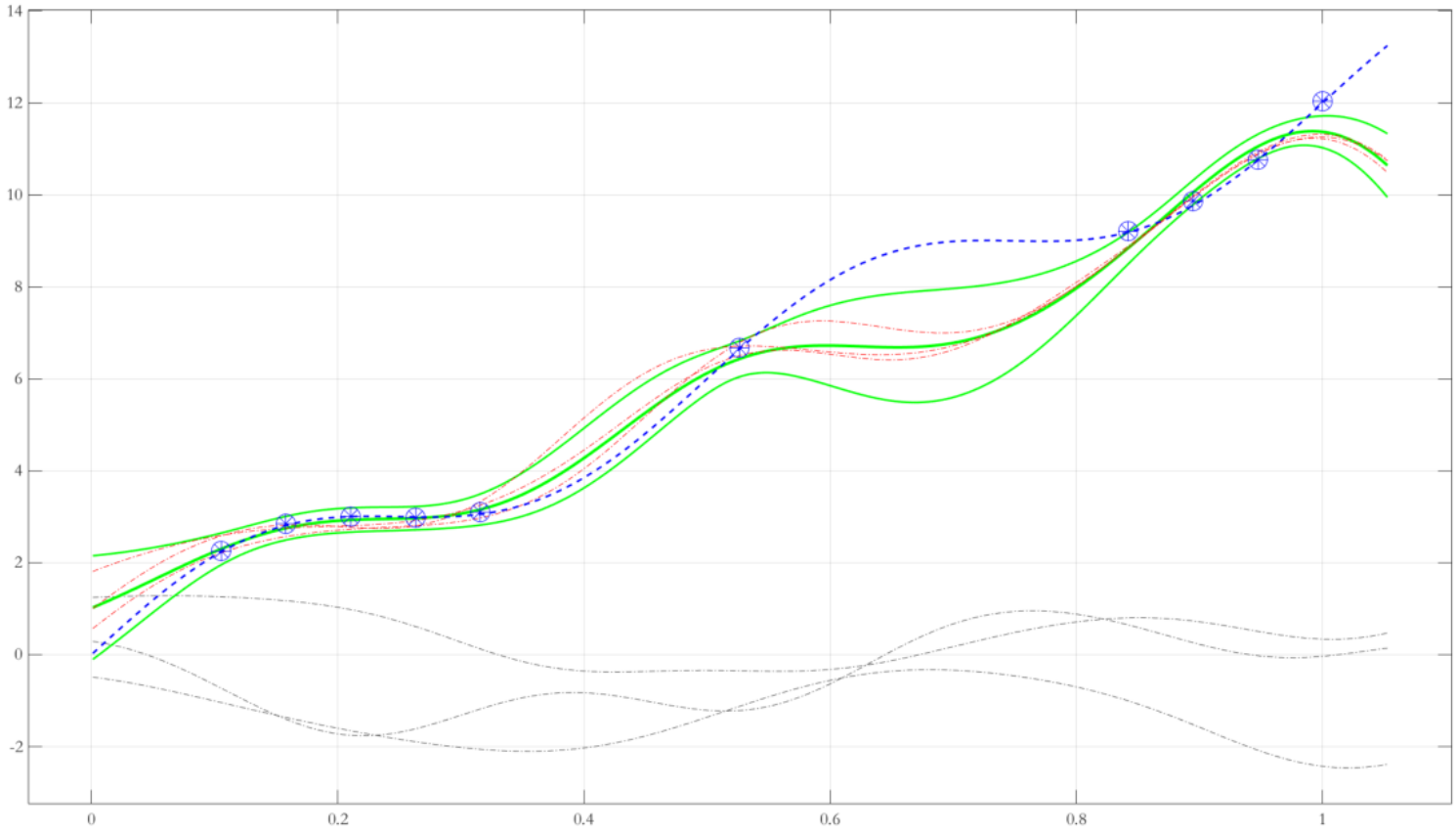
Glas = gwir; Gwyrdd = amcangyfrif o gasgliad Bayesaidd
$y = \mathcal{GP}(x|\lambda)$, model proses Gaussaidd â hyper-baramedr λ sydd yn rhy fach
Llwyd= efelychiadau o dan y rhag-fodel; Coch = efelychiadau o dan yr ôl-fodel



Rhy arw!

Gwyddor data, rhyngosod ac allosod

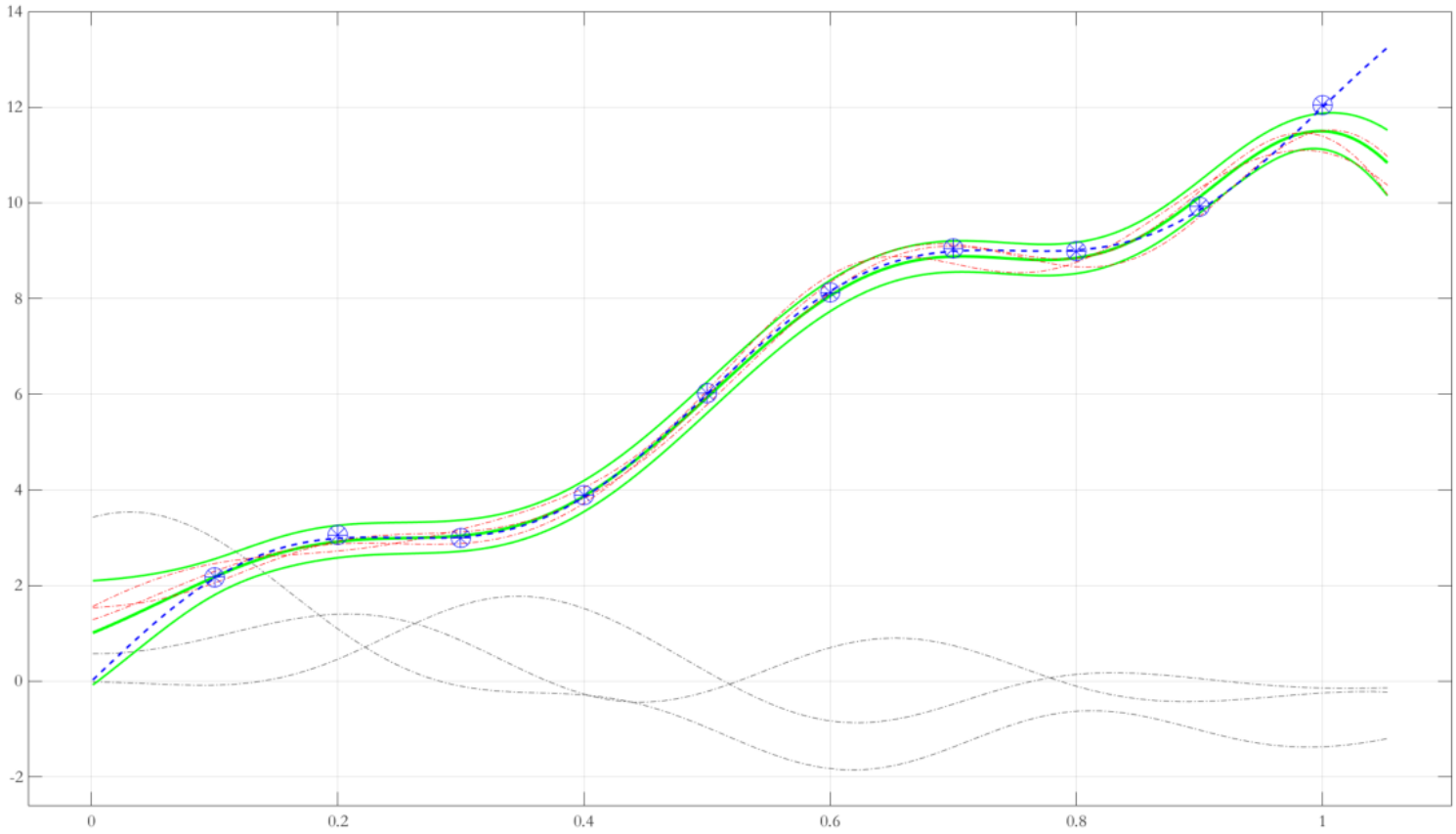
Glas = gwir; Gwyrdd = amcangyfrif o gasgliad Bayesaidd
$y = \mathcal{GP}(x|\lambda)$, model proses Gaussaidd â hyper-baramedr λ sydd yn rhesymol
Llwyd= efelychiadau o dan y rhag-fodel; Coch = efelychiadau o dan yr ôl-fodel



Da!

Gwyddor data, rhyngosod ac allosod

Glas = gwir; Gwyrdd = amcangyfrif o gasgliad Bayesaidd
$y = \mathcal{GP}(x|\lambda)$, model proses Gaussaidd â dewisiadau ar gyfer x sydd yn gofod-lenwi
Llwyd= efelychiadau o dan y rhag-fodel; Coch = efelychiadau o dan yr ôl-fodel



Gwell!

Gwyddor data

Cymwysiad call

- Iawn os ydym yn sicr ein bod yn "rhyngosod"
- Y broblem yw allosod
- Sut mae asesu rhyngosod ac allosod mewn gofod 1000-ddimensiynol?
- Penu ffwythiant cnewyllyn a hyper-baramedrau yn ganolog
- Syniadau ystadegol dylunio arbrawf yn hanfodol
- Asesu manwl o ansawdd rhagfynegiadau

Ond

- Nid yw rheolau mathemateg wedi newid!
- Nid oes gwyddoniaeth newydd yma, tan i ni ddechrau meddwl am gyfrifo cwantwm, neu ddeallusrwydd **cyffredinol** artiffisial

Gwyddor data

Am beth mae'r ffŷs a'r ffwdan?

- **Cyd-gyfeiriad meysydd gwahanol**
- Casglu data ar raddfa anhygoel
- Bron popeth (mae'n ymddangos) "ar-lein" wedi ei ddigido
- Rhwydweithiau cyd-gysylltiedig, cyflym, byd-eang
- Gwellianau cyfrifiadurol (cyflymder, cof, cost, "y cwmwl")
- Algorithmau slic i amcangyfrif modelau empirig
- Y rhyngwyneb rhwng dyn a chyfrifiadur yn aeddfedu

Felly, yn ymarferol

- Mynediad hawdd i bob data o ddiddordeb o bobman
- Mynediad hawdd i feddalwedd ystadegol safonol hawdd-ei-defnyddio rhad-ac-am-ddim (e.e. R, PYTHON)
- Mynediad hawdd i gyfrifiaduron pwerus

Problem priodi

Dychmygwch

- ’Da chi’n ddyn sengl yn edrych am wraig
- ’Da chi’n mynd i gwrdd â 100 o ferched cymwys ar hap dros nifer fawr o flynyddoedd
- Byddwch chi’n hoffi rhai’n well nag eraill, ac yn hoffi un fwyaf oll!
- Mae’n rhaid i chi benderfynu’n syth pan ’da chi’n cwrdd â merch, os taw hi **yw’r un**
- ’Da chi ddim eisiau aros am byth i briodi, wrth rheswm!

- Sawl merch dylsech chi gwrdd â nhw cyn penderfynu?

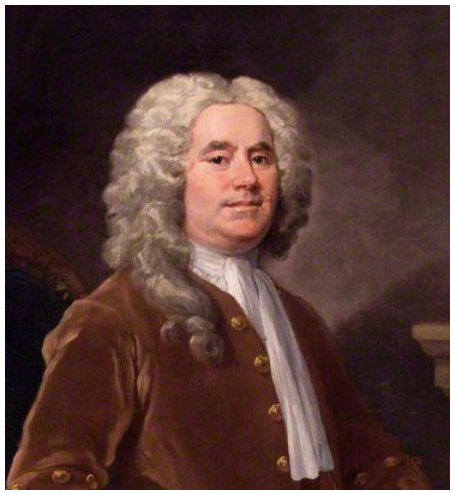
- Cynigion?

Y problem briodas

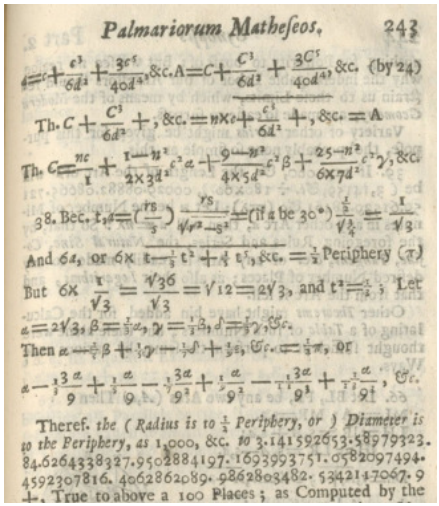
- Mae'r ateb yn esiampl o reol stopio optimaidd, un o gonglfeini maes a elwir yn ymchwil gweithredol, a dyma fe ateb
- Dylsech chi aros tan i chi gwrdd â $100/e \approx 37$ o ferched a chofio o'r rheini pa un oedd orau (dywedwn ni taw Carys oedd enw hon)
- Wedyn dylsech chi ddal i gwrdd â merched, a dewis fel gwraig y ferch gyntaf 'da chi'n cwrdd â hi sy'n well na Carys
- Fel hyn, mae gyda chi'r tebygolrwydd uchaf o gwrdd â'r gorau o'r 100, a'r tebygolrwydd hwnw yw $1/e \approx 0.37$
- e yw cysonyn Euler, $e = 2.7182818...$

Hafaliad prydferthaf mathemateg

$$e^{i\pi}+1=0$$



William Jones
(Llanfihangel Tre'r Beirdd, 1675 – 1749)


Palmariorum Matheſeos, 243

Llyfr William Jones

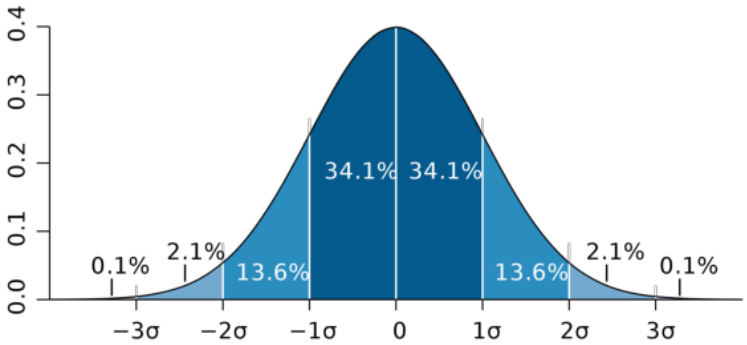
Robert Recorde
(Dinbych-y-pysgod, 1512–1558)

Dosraniadau'r cyffredin ... a'r anghyffredin

Terfan ganolog $\Rightarrow$ **dosraniad Gaussaidd neu "normal"** (Lindeberg-Lévy)

$$\lim_{n\to\infty} \mathbb{P}[\sqrt{n}(\frac{X_1 + X_2 + ... + X_n}{n} - \mu) \leq z] = \Phi(z/\sigma)$$

$$\phi(z/\sigma) = \frac{d\Phi}{dz} = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{z^2}{2\sigma^2})$$



Dwysedd Gaussian (o wiki)

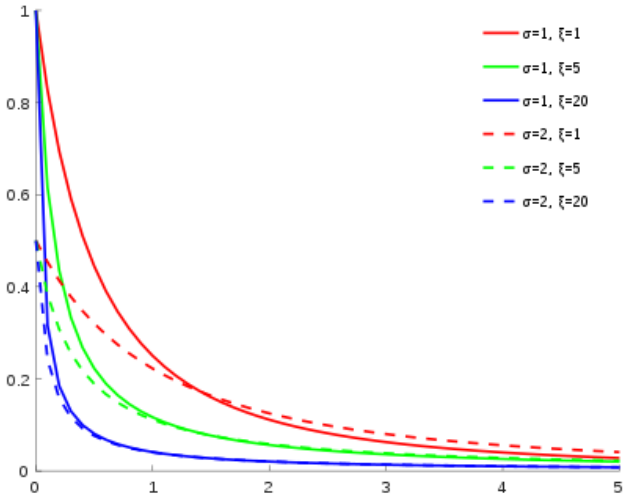
Rhydd ϕ ffurf ar gyfer tebygoliaeth sampl mewn casgliad Bayesaidd.

Dosraniadau'r cyffredin ... a'r anghyffredin

Brigau dros trothwy $\Rightarrow$ **dosraniad Pareto cyffredinoledig** (Pickands-Balkema-De Haan)

$$\lim_{\psi \to \infty} \mathbb{P}[X \leq z | X > \psi] = \mathrm{GP}(z|\xi, \sigma, \psi)$$

$$\mathrm{GP}(z|\xi, \sigma, \psi) = (1 + \frac{\xi}{\sigma}(x - \psi))^{-1/\xi}$$



Dwysedd GP (o wiki)

Rhydd GP debygoliaeth sampl o eithafon mewn casgliad Bayesaidd.

Meintioli risg

- System â phriodweddau $\mathcal{S}$ yn creu allbynau Y o fewnbynau X
- Budd $U(Y|\mathcal{S})$ (neu gost, risg) y system yn ddibynol ar Y ac $\mathcal{S}$
- X yn dibynu ar gyd-newidynau Θ
- Pob un o Θ, X, Y yn hap-newidyn (aml-ddimensiynol)

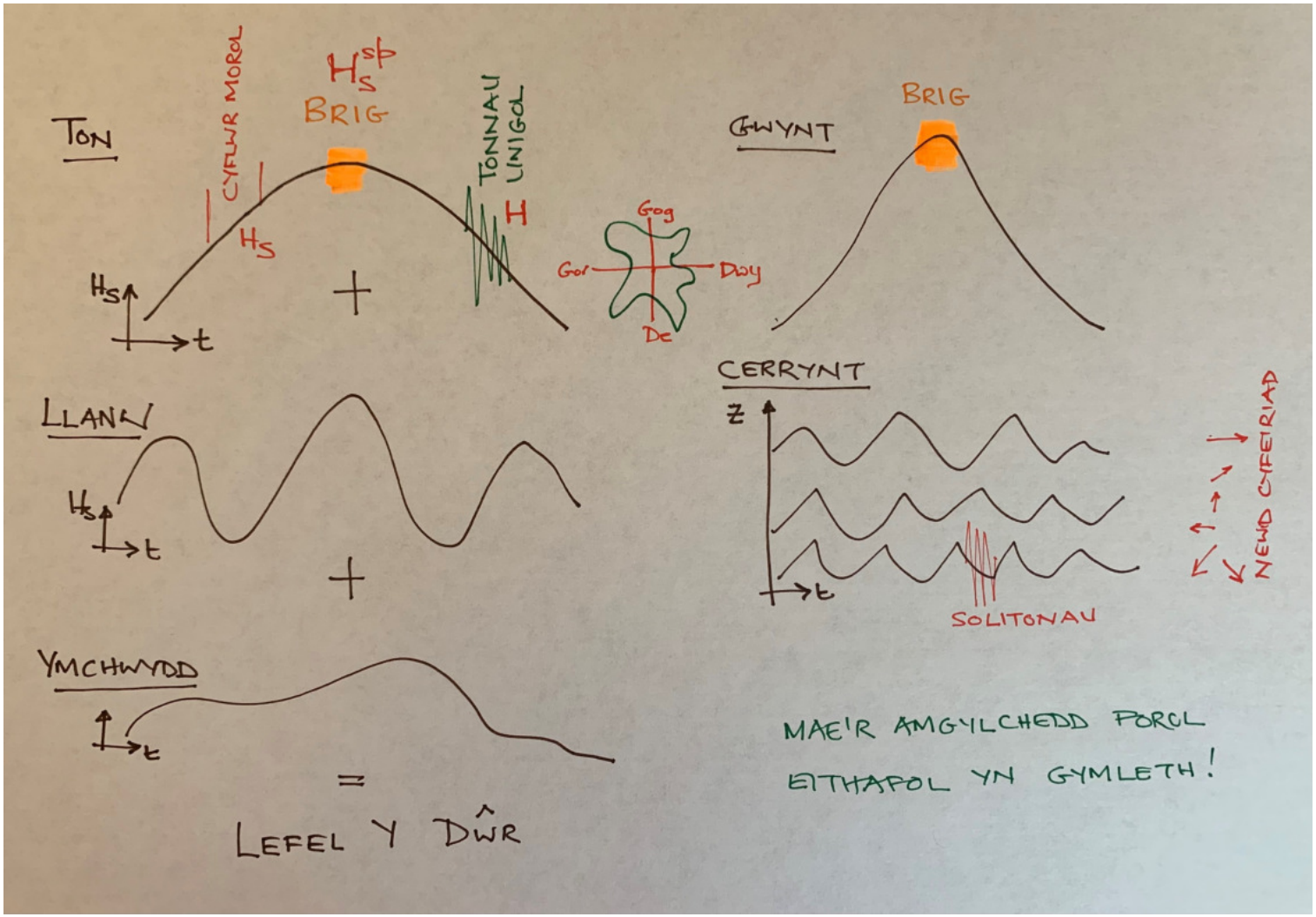
Sut gallwn feintioli budd y system, a newid $\mathcal{S}$ i uchafsymio'r budd?

- Amcangyfrif model $f_{Y|X,\Theta}$ am ddwysedd yr allbwn
- Amcanfyfrif model $f_{X|\Theta}$ am ddwysedd y mewnbwn
- Amcangyfrif model f_{Θ} am ddwysedd y cyd-newidynau

$$\mathbb{E}[U|\mathcal{S}] = \int_y \int_x \int_\theta U(y|\mathcal{S}) f_{Y|X,\Theta}(y|x,\theta) f_{X|\Theta}(x|\theta) f_{\Theta}(\theta) d\theta dx dy$$

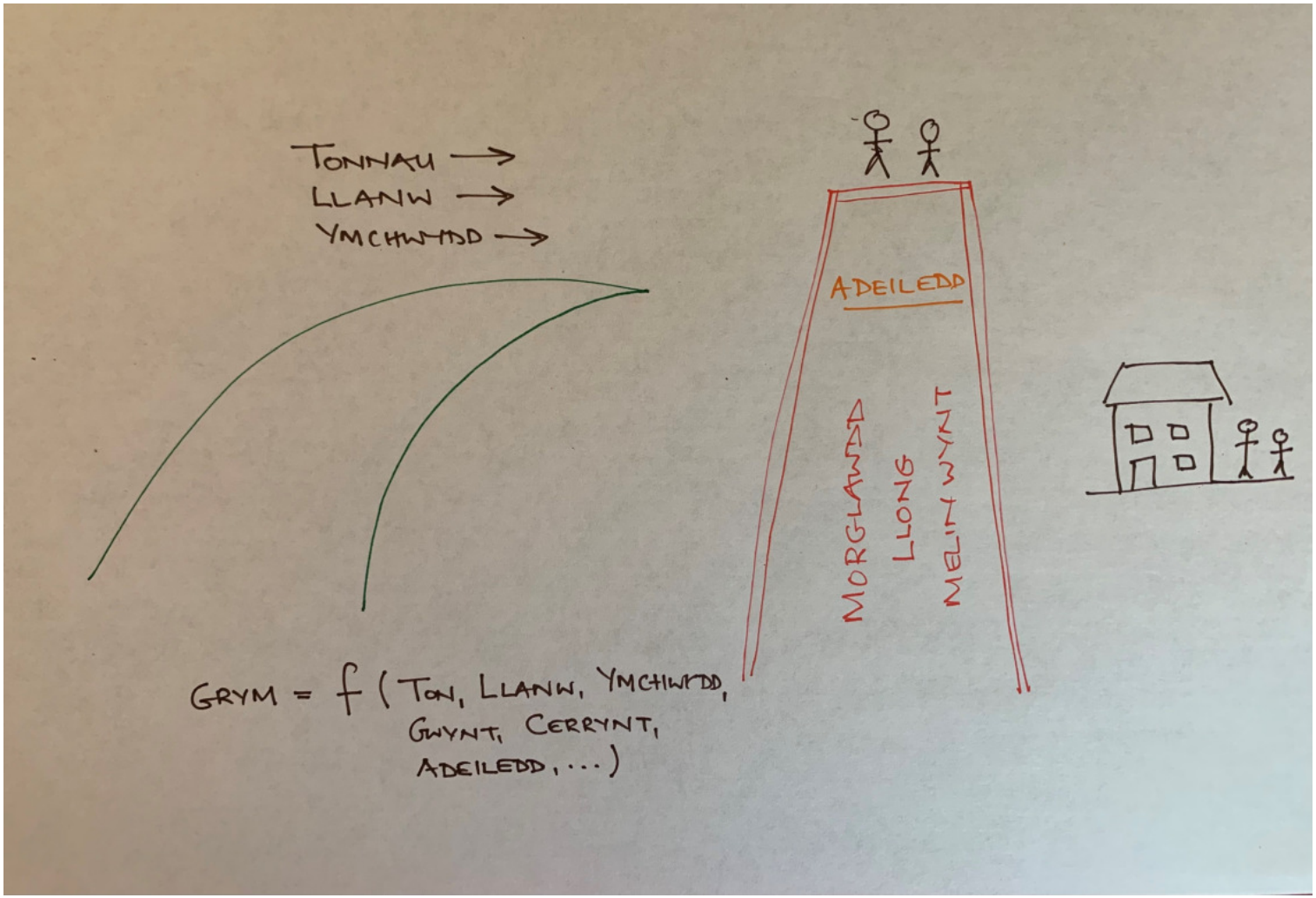
$\Rightarrow$ Gallwn newid $\mathcal{S}$ i uchafsymio'r budd!

Yr amgylchedd forol eithafol



... mae'n gymleth

Yr amgylchedd forol eithafol



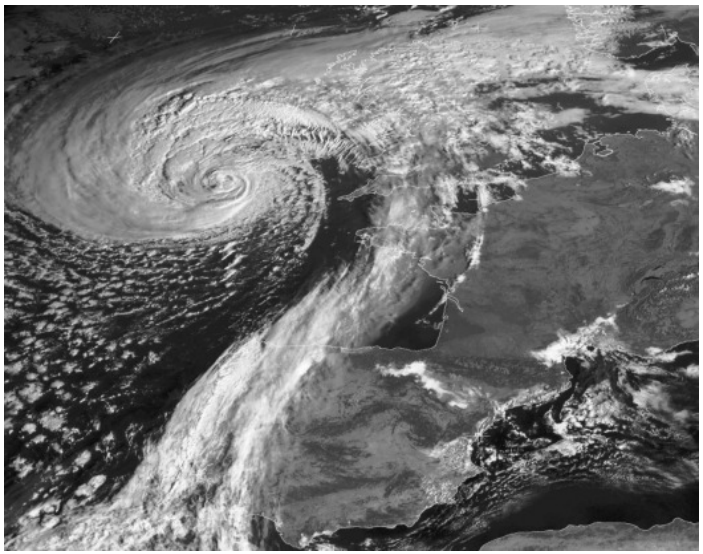
... mae'n gymleth iawn

Dibynadwyedd adeileddau morol ac arfordirol

- **Dibynadwyedd**: beth yw'r tebygolrwydd y bydd adeiledd morol cryfder S yn methu mewn N mlynedd?

$$\int_L \int_H I(L > S) f(L|H) \boldsymbol{f_N(H)} dH dL$$

- Creir llwyth L gan **donnau**, gwyntoedd a cheryntau
- H yw uchder y don
- **Bwriad**: amcangyfrif dosbarthiad hir-dymor H



Storm Ophelia (2017)

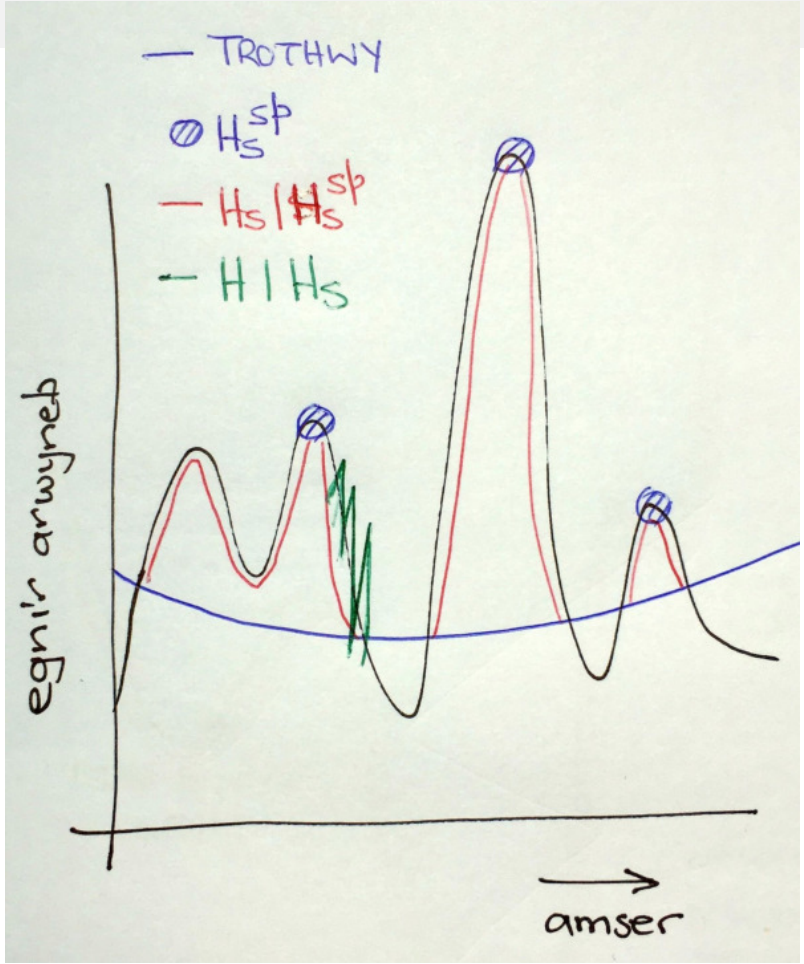
Cormorant Alpha

Model hierarchaidd

- **Mae angen $f(H)$** hir-dymor
- Ond rhaid modelu H, H_S a H_S^{sp}

- ***H***: uchder ton unigol cafn-brig
- H_S: egni'r arwyneb dros awr
- H_S^{sp}: egni brig y storom

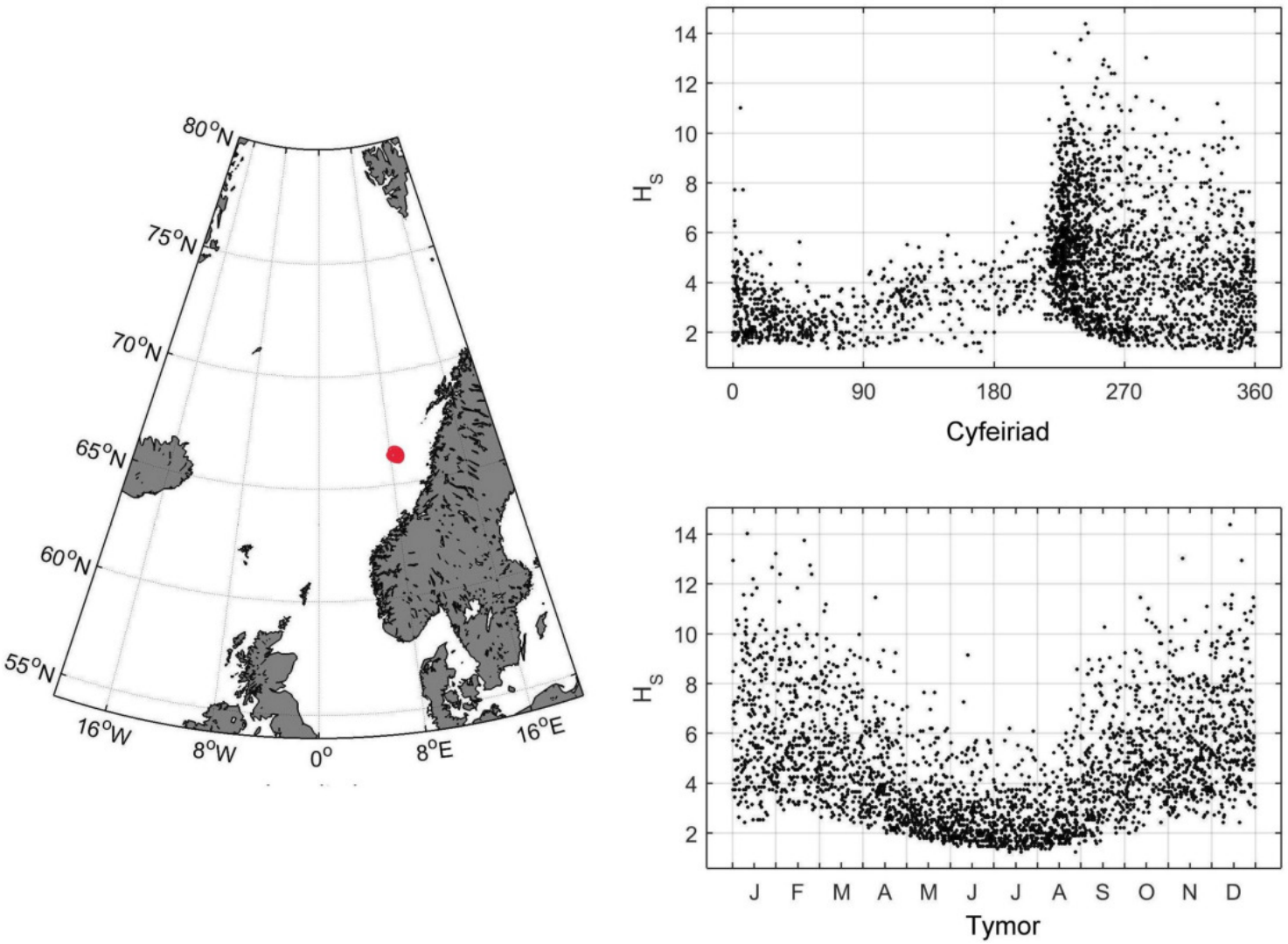
- $H|H_S$: H byr-tymor
- $H_S|H_S^{sp}$: esblygiad y storom
- H_S^{sp}: egni brig **hir-dymor**



Storm forol yn syml

- Cyd-ddosbarthiad: $f(H, H_S, H_S^{sp}) = f(H|H_S)f(H_S|H_S^{sp})f(H_S^{sp})$
- Yn ymylol: $f(H) = \int_{H_S} \int_{H_S^{sp}} f(H|H_S)f(H_S|H_S^{sp})\boldsymbol{f(H_S^{sp})}\, dH_S dH_S^{sp}$

Cymhwysiad Môr Y Gogledd: y data

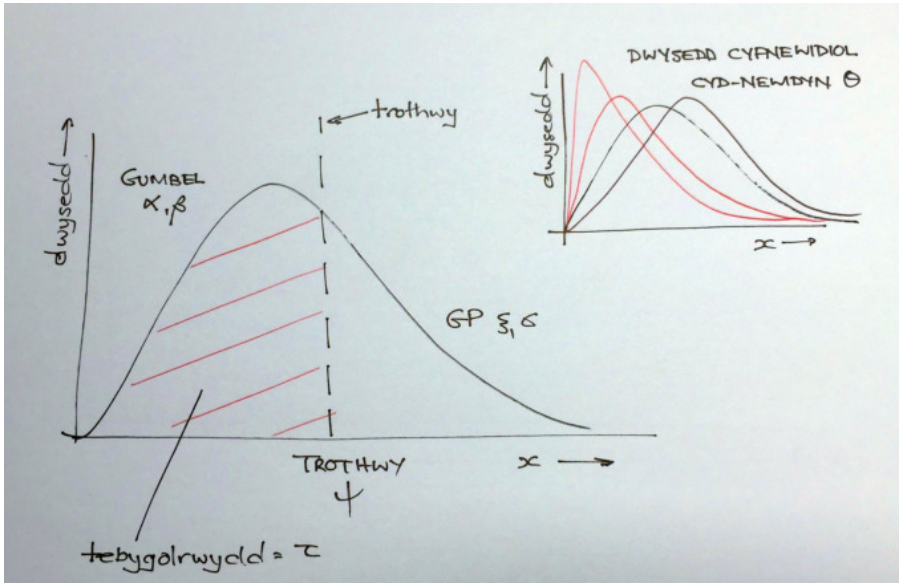


Digwyddiadau H_S o Fôr Y Gogledd gyda chyfeiriad a thymor

Modelu eithafon

- Damcaniaeth asymptotig brigau dros drothwy
- Dosbarthiad Pareto cyffredinoledig (**GP**)
  $f_{X|X>\psi}(x|\psi) = \frac{1}{\sigma}(1 + \frac{\xi}{\sigma}(x-\psi))_+^{1/\xi - 1}, \quad x, \psi, \xi \in \mathbb{R}, \sigma \in \mathbb{R}^{>0}$
- e.e. $X = H_S^{sp}$
- Casgliad **Bayesaidd**

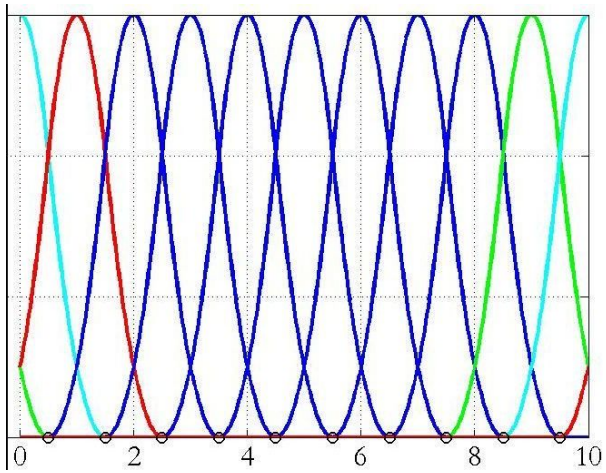
- Effeithiau cyd-newidynau
- $\eta \in \{\psi, \sigma, \xi\}, \quad \eta \equiv \eta(\theta_1, \theta_2, \ldots)$
- e.e. θ = cyfeiriad y storom, tymor, dyfnder, cyrch



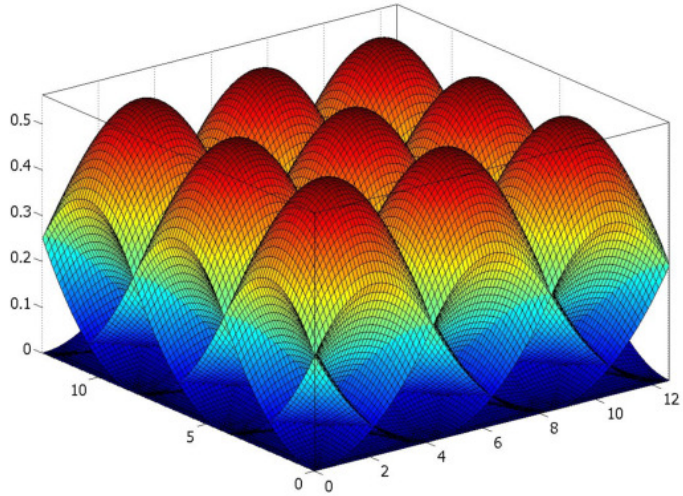
Dwysedd Gumbel blaendor-GP

Cynrychioliad sail amharametrig

- Modd **hyblyg** i ddisgrifio amrywiad $\eta \in \{\psi, \sigma, \xi\}$ â $\theta_1, \theta_2, \ldots$
- Seiliau **splein** $\boldsymbol{\eta} = \boldsymbol{B}\boldsymbol{\beta}_\eta$ aml-ddimensiynol $\boldsymbol{B} = \boldsymbol{B}_{\theta_1} \otimes \boldsymbol{B}_{\theta_2} \otimes \ldots$
- Seiliau **Voronoi** / **brithwaith**
- Cosbi **garwedd** $\lambda_\eta \boldsymbol{\beta}'_\eta \boldsymbol{P} \boldsymbol{\beta}_\eta$ â matrics cosb $\boldsymbol{P}$ a chyfeirnod λ_η



Sail splein cyfnodol 1-D

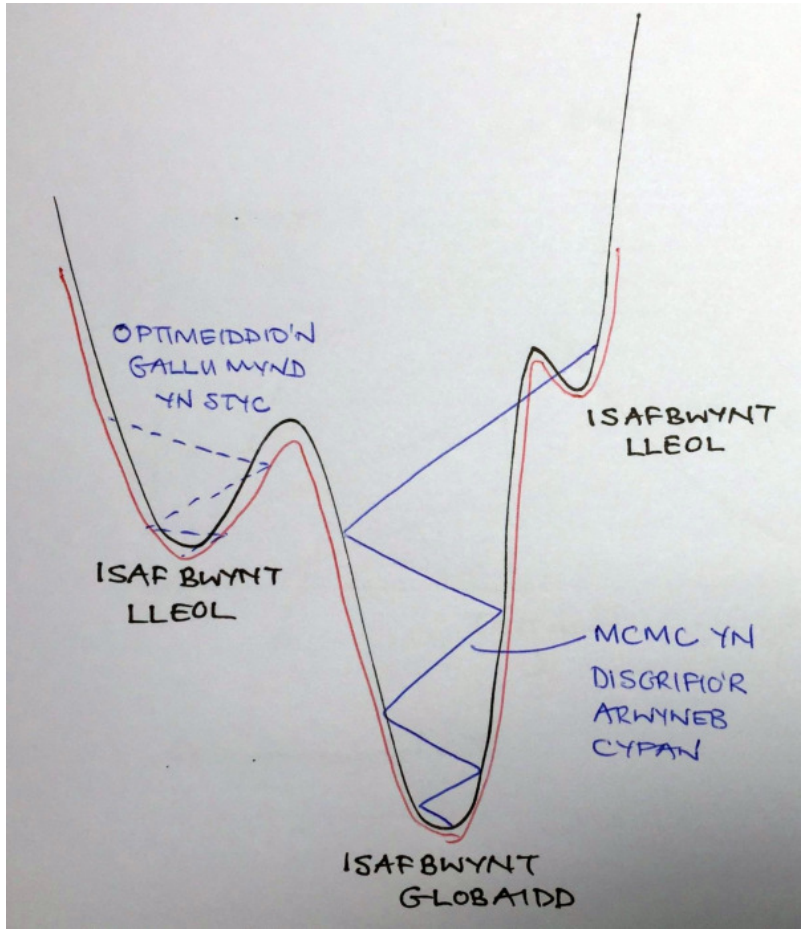


Sail splein lluoswm tensor 2-D

Casgliad Bayesaidd

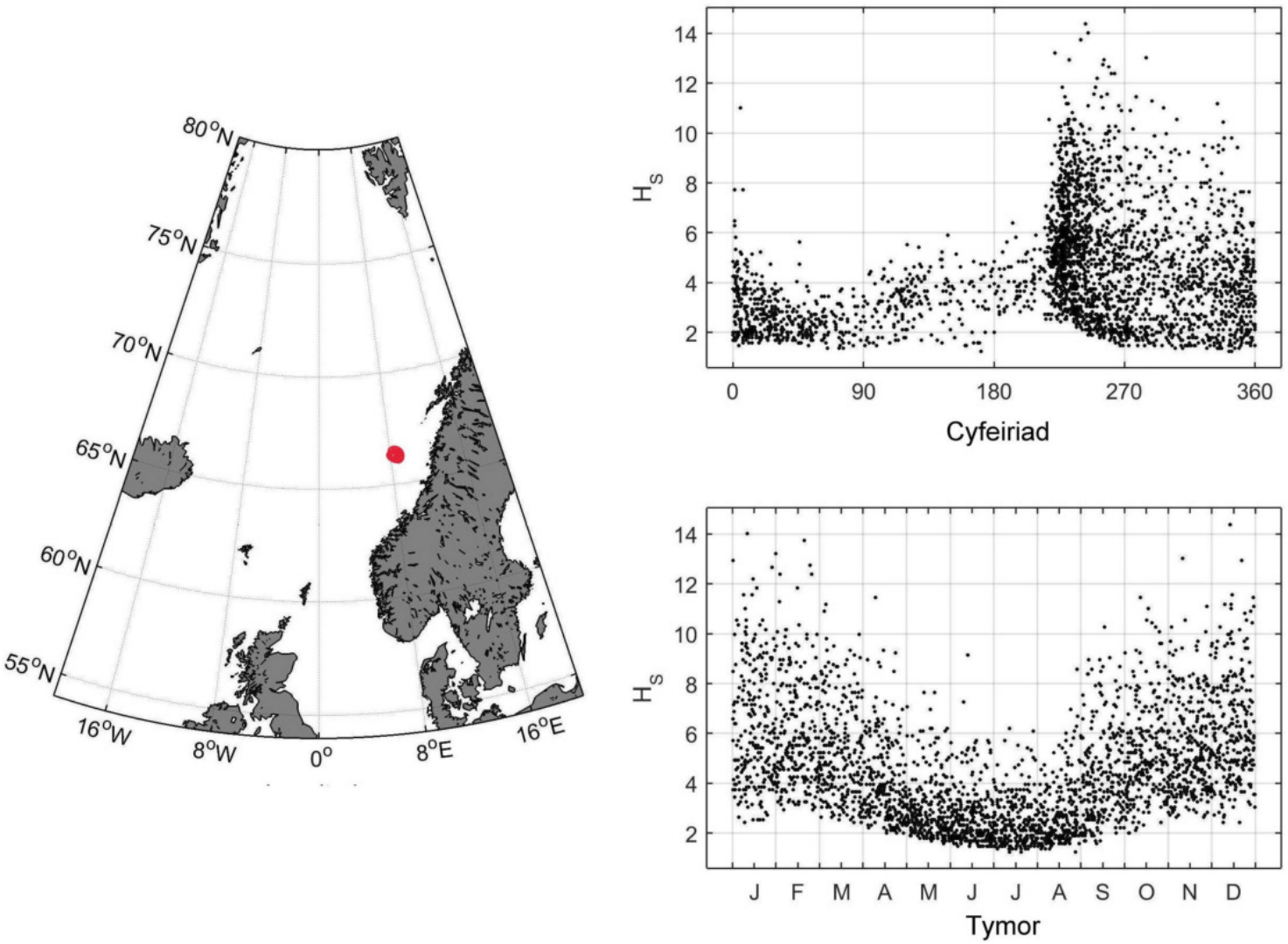
Bwriad: dysgu am $\boldsymbol{\beta}_\eta, \lambda_\eta, \eta \in \{\psi, \sigma, \xi\}$
Yn syml, ar gyfer paramedrau ζ:

- Penu **rhag-ddwysedd**
  $f(\zeta) \propto \exp(-\lambda_\eta \boldsymbol{\beta}_\eta' \boldsymbol{P} \boldsymbol{\beta}_\eta)$
- Theorem Bayes / MCMC:
  $f(\zeta|X) \propto f(X|\zeta) f(\zeta)$
- Samplu amodol **iterus**:
  $f(\zeta_1|X, \zeta_2) \propto f(X|\zeta_1, \zeta_2) f(\zeta_2)$
  $f(\zeta_2|X, \zeta_1) \propto f(X|\zeta_1, \zeta_2) f(\zeta_1)$
  Metropolis-Hastings
  mMALA
- Amcangyfrif **ôl-ddwysedd**
  $f(\zeta|X)$



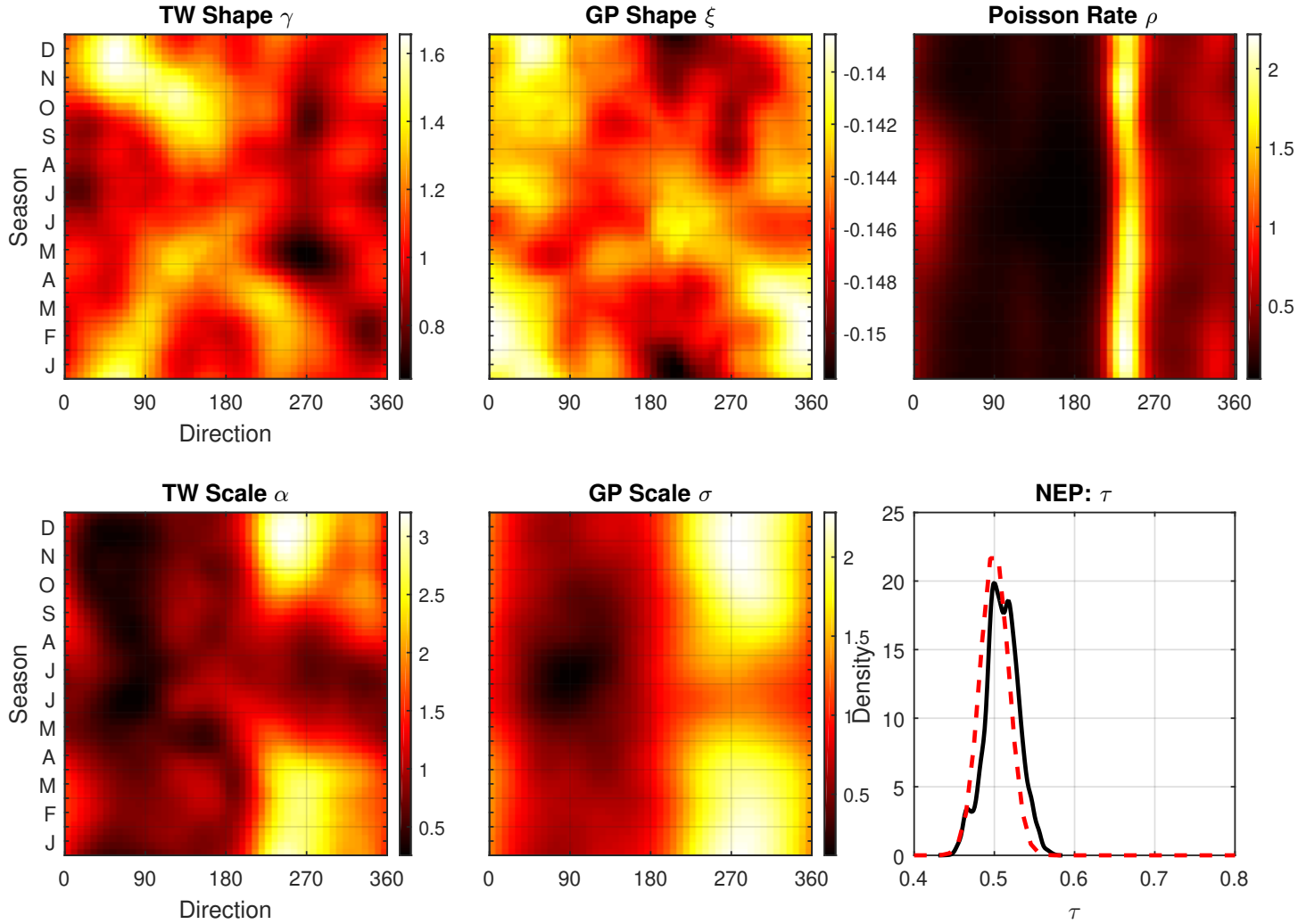
Casgliad Bayesaidd yn syml

Cymhwysiad Môr Y Gogledd: y data



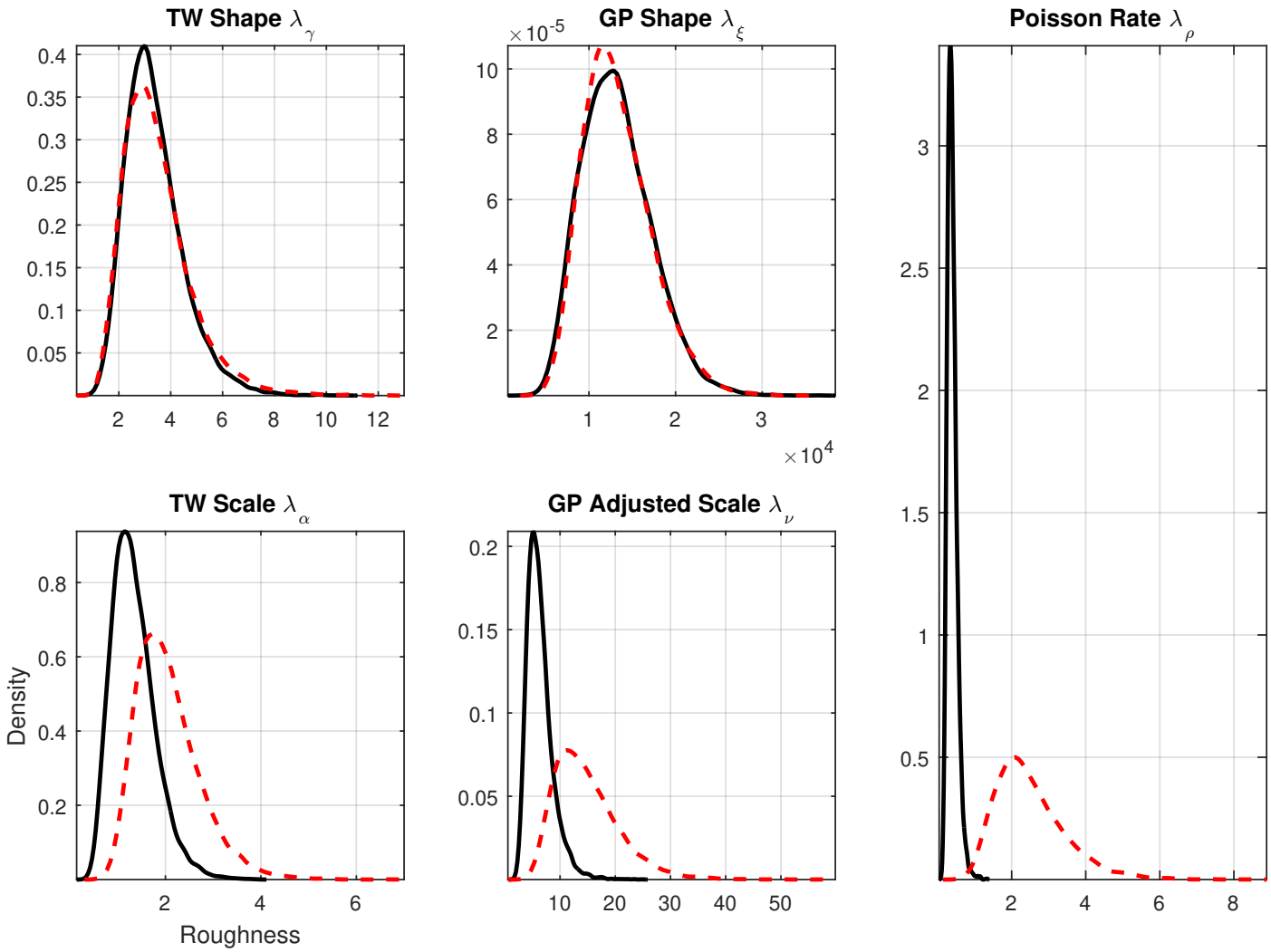
Digwyddiadau H_S o Fôr Y Gogledd gyda chyfeiriad a thymor

Amcangyfrifon paramedrau



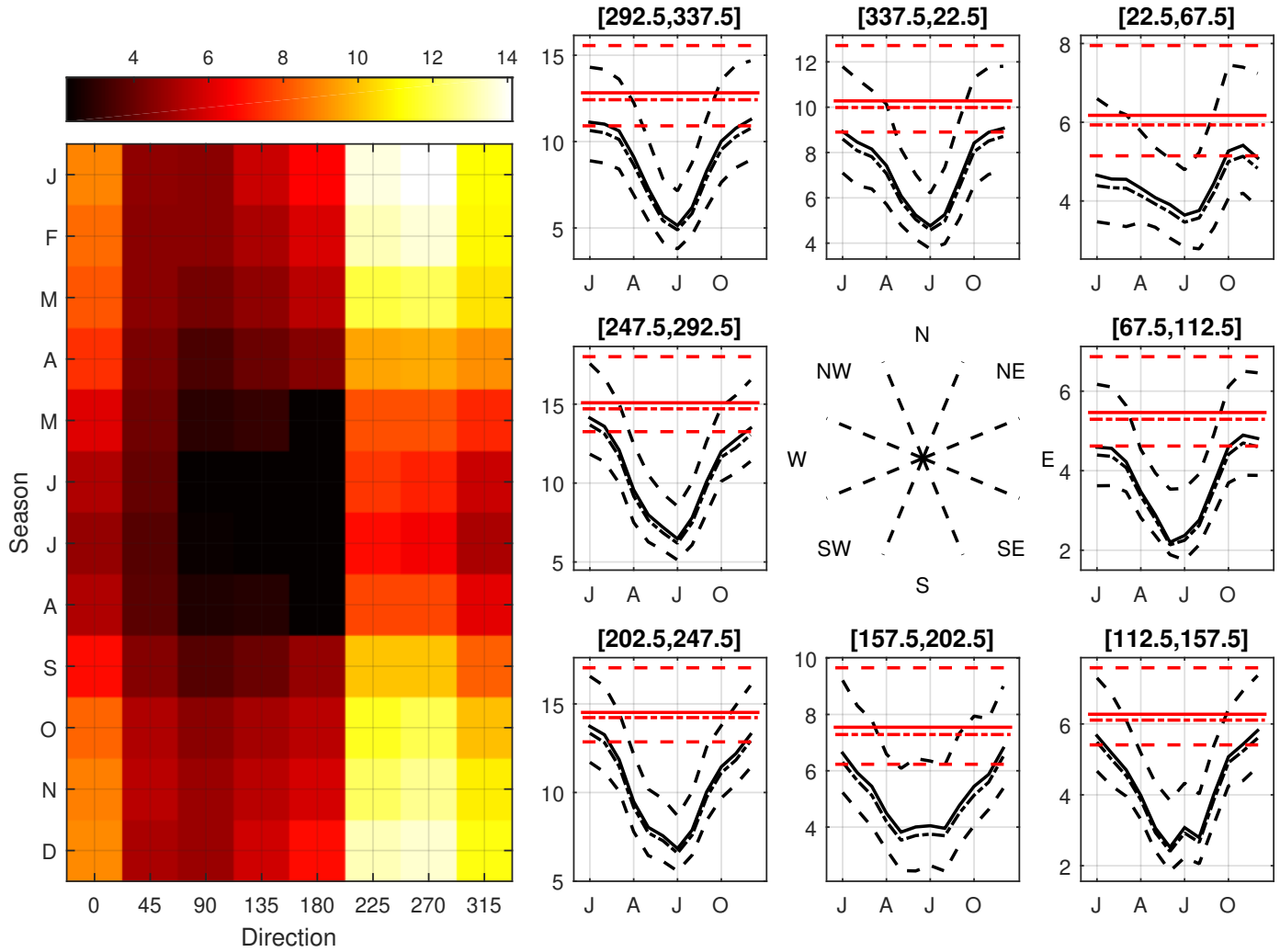
Ôl-ganolrifau ξ a σ (yn y golofn ganol) ar gyfer model H_S^{sp}

Amcangyfrifon cyfeirnodau garwedd



Dwyseddau cyfeirnodau cosb ξ a σ (yn y golofn ganol) ar gyfer model H_S^{sp}; Randell et al. 2016

Gwerth mwyaf *N*-mlynedd



Ôl-ganolrifau gwerth mwyaf 100-mlynedd H_S^{sp} â chyfeiriad a thymor (y chwith); amrywiad misol â chyfeiriad (y dde)

Estyniadau a chymwysiadau pellach

Eithafon **aml-newidynol**

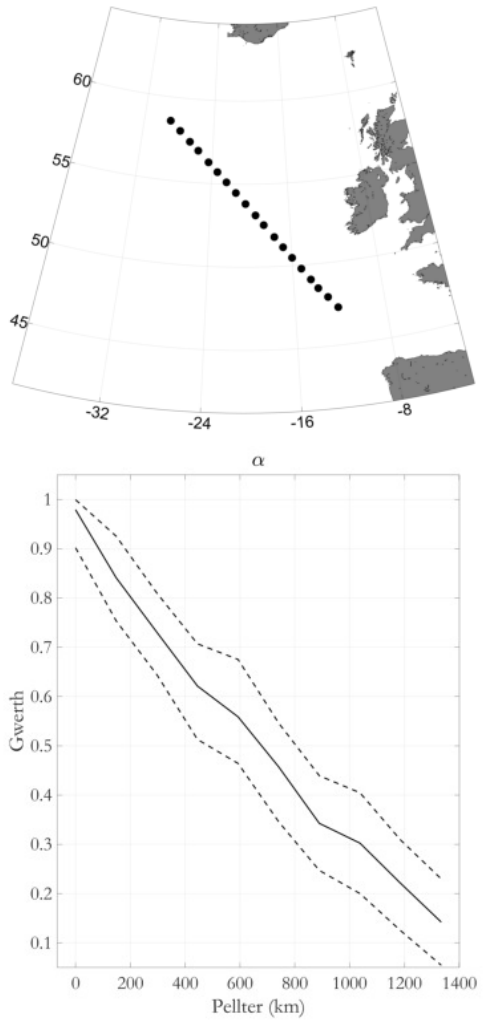
- e.e. cyd-ddosbarthiad tonnau, gwyntoedd a cheryntau

Eithafon **gofodol-amserol**

- e.e. $H_S(r_1, r_2, ...)|\{H_S(r_0) > \psi\}$

Goblygiadau amgylcheddol

- gwell asesiad o ddibynadwyedd



Dibyniaeth eithafol ofodol H_S o ddata lloerenau JASON, damcaniaeth eithafon gofodol amodol

Diolch yn fawr ☺

(www.lancs.ac.uk/∼jonathan)