Supporting Information

Censored regression modelling to predict virus inactivation in wastewaters, *Environmental Science & Technology*

Dr. Julii Brainard[1], Dr. Katherine Pond[2], Professor Paul R. Hunter[1]*

[1] Norwich Medical School, University of East Anglia Norwich NR4 7TJ, United Kingdom

[2] Department of Civil and Environmental Engineering, Robens Centre for Public and Environmental Health, University of Surrey, Guildford GU2 7XH

* Professor Paul Hunter, Norwich Medical School, University of East Anglia Norwich NR4 7TJ, United Kingdom. Tel. 01603-456161. Fax. 01603-593752. Paul.Hunter@uea.ac.uk.

Content Summary

Number of pages: 11
Number of figures: 2
Number of tables: 7

Section S1. References used to record virus attributes: family, genus, RNA or DNA virus, enveloped or not, and primary transmission pathway. [1-47]

1.	Abate, E.; Belayneh, M.; Idh, J.; Diro, E.; Elias, D.; Britton, S.; Aseffa, A.; Stendahl, O.; Schön, T., Asymptomatic helminth infection in active tuberculosis is associated with increased regulatory and th-2 responses and a lower sputum smear positivity. *PLoS Negl. Trop. Dis.* **2015,** *9* (8).

2.	Bae, J.; Schwab, K. J., Evaluation of murine norovirus, feline calicivirus, poliovirus, and MS2 as surrogates for human norovirus in a model of viral persistence in surface water and groundwater. *Appl. Environ. Microbiol.* **2008,** *74* (2), 477-484.

3.	Bøtner, A.; Belsham, G. J., Virus survival in slurry: analysis of the stability of foot-and-mouth disease, classical swine fever, bovine viral diarrhoea and swine influenza viruses. *Vet. Microbiol.* **2012,** *157* (1), 41-49.

4.	Bovilis, Bovine Rotavirus. **undated**, http://www.bovilis.com/diseases/ncd/causes/bovine-rotavirus.asp (May 24).

5.	Broad Institute, Herpes Simplex Virus (HSV) Classification. In *Herpes Simplex Virus Portal*, Undated.

6.	Casanova, L.; Rutala, W. A.; Weber, D. J.; Sobsey, M. D., Survival of surrogate coronaviruses in water. *Water Res.* **2009,** *43* (7), 1893-1898.

7.	Casjens, S., Information about bacteriophage P22. **2000**, https://www.asm.org/division/m/fax/P22Fax.html (2016-07-01).

8.	Center for Food Security and Public Health, Aujesky's Disease. **2006**, http://www.cfsph.iastate.edu/Factsheets/pdfs/aujeszkys_disease.pdf (2016-07-01).

9.	Centre for Agriculture and Biosciences International, Bovine enterovirus infection. **2010**, http://www.cabi.org/isc/datasheet/91707 (2016-07-01).

10.	Cole, D.; Long, S. C.; Sobsey, M. D., Evaluation of F+ RNA and DNA coliphages as source-specific indicators of fecal contamination in surface waters. *Appl. Environ. Microbiol.* **2003,** *69* (11), 6507-6514.

11.	Comparative Medicine Program, Mouse Hepatitis Virus (MHV). **2013**, http://dora.missouri.edu/mouse/mouse-hepatitis-virus-mhv/ (2016-07-01).

12.	Daugelavičius, R.; Cvirkaitė, V.; Gaidelytė, A.; Bakienė, E.; Gabrėnaitė-Verkhovskaya, R.; Bamford, D. H., Penetration of enveloped double-stranded RNA bacteriophages φ13 and φ6 into Pseudomonas syringae cells. *J. Virol.* **2005,** *79* (8), 5017-5026.

13.	ESR Ltd, Enteric Viruses. **2001**, http://www.foodsafety.govt.nz/elibrary/industry/Enteric_Viruses-Other_Than.pdf (2016-07-01).

14.	Elving, J.; Emmoth, E.; Albihn, A.; Vinnerås, B.; Ottoson, J., Composting for avian influenza virus elimination. *Appl. Environ. Microbiol.* **2012,** *78* (9), 3280-3285.

15.	Emmoth, E.; Ottoson, J.; Albihn, A.; Belák, S.; Vinnerås, B., Ammonia disinfection of hatchery waste for elimination of single-stranded RNA viruses. *Appl. Environ. Microbiol.* **2011,** *77* (12), 3960-3966.

16.	Ghebremedhin, B., Human adenovirus: Viral pathogen with increasing importance. *Eur. J. Microbiol. Immunol. (Bp)* **2014,** *4* (1), 26-33.

17.	Geoghegan, J. L.; Senior, A. M.; Giallonardo, F. D.; Holmes, E. C., Virological factors that increase the transmissibility of emerging human viruses. *Proc. Natl. Acad. Sci. U. S. A.* **2016,** *113* (15), 4170-4175.

18.	Hardestam, J.; Simon, M.; Hedlund, K.; Vaheri, A.; Klingström, J.; Lundkvist, Å., Ex vivo stability of the rodent-borne Hantaan virus in comparison to that of arthropod-borne members of the Bunyaviridae family. *Appl. Environ. Microbiol.* **2007,** *73* (8), 2547-2551.

19.	Karst, S. M., Pathogenesis of noroviruses, emerging RNA viruses. *Viruses* **2010,** *2* (3), 748-781.

20.	Kuhn, J. H.; Becker, S.; Ebihara, H.; Geisbert, T. W.; Johnson, K. M.; Kawaoka, Y.; Lipkin, W. I.; Negredo, A. I.; Netesov, S. V.; Nichol, S. T., Proposal for a revised taxonomy of the family Filoviridae: classification, names of taxa and viruses, and virus abbreviations. *Arch. Virol.* **2010,** *155* (12), 2083-2103.

21.	Levy, J.; Hutsell, S., Overview of Feline Infectious Peritonitis. **2014**, http://www.merckvetmanual.com/mvm/generalized_conditions/feline_infectious_peritonitis/overview_of_feline_infectious_peritonitis.html (2016-07-01).

22.	Lewis, C. S.; Porter, E.; Matthews, D.; Kipar, A.; Tasker, S.; Helps, C. R.; Siddell, S. G., Genotyping coronaviruses associated with feline infectious peritonitis. *J. Gen. Virol.* **2015,** *96* (6), 1358-1368.

23.	Lin, T.-Y.; Lo, Y.-H.; Tseng, P.-W.; Chang, S.-F.; Lin, Y.-T.; Chen, T.-S., A T3 and T7 recombinant phage acquires efficient adsorption and a broader host range. *PLoS One* **2012,** *7* (2), e30954.

24.   Magri, M. E.; Fidjeland, J.; Jönsson, H.; Albihn, A.; Vinnerås, B., Inactivation of adenovirus, reovirus and bacteriophages in fecal sludge by pH and ammonia. *Sci. Total Environ.* **2015**, *520*, 213-221.

25.   Marrero, R.; Limardo, R. R.; Carrillo, E.; König, G. A.; Turjanski, A. G., A computational study of the interaction of the foot and mouth disease virus VP1 with monoclonal antibodies. *J. Immunol. Methods* **2015**, *425*, 51-57.

26.   Mayer, G., Bacteriophage. In *Microbiology and Immunology Online*, Hunt, R., Ed. **2016**.

27.   McGrath, S.; van Sinderen, D., *Bacteriophage: Genetics and Molecular Biology*. Horizon Scientific Press: **2007**.

28.   Moser, C.; Stettler, P.; Tratschin, J.-D.; Hofmann, M. A., Cytopathogenic and noncytopathogenic RNA replicons of classical swine fever virus. *J. Virol.* **1999**, *73* (9), 7787-7794.

29.   Nielsen, A. A.; Skovgård, H.; Stockmarr, A.; Handberg, K. J.; Jørgensen, P. H., Persistence of low-pathogenic avian influenza H5N7 and H7N1 subtypes in house flies (Diptera: Muscidae). *J. Med. Entomol.* **2011**, *48* (3), 608-614.

30.   Public Health Agency of Canada, Echovirus - Pathogen Safety Data Sheet. **2011**, http://www.phac-aspc.gc.ca/lab-bio/res/psds-ftss/msds56e-eng.php (2016-5-24).

31.   Puig, M.; Girones, R., Genomic structure of phage B40-8 of Bacteroides fragilis. *Microbiology* **1999**, *145* (7), 1661-1670.

32.   Rabenau, H.; Cinatl, J.; Morgenstern, B.; Bauer, G.; Preiser, W.; Doerr, H., Stability and inactivation of SARS coronavirus. *Med. Microbiol. Immunol.* **2005**, *194* (1-2), 1-6.

33.   Rad, L. H. Removal of Salmonella Typhimurium phage 28B by nanoscale zero-valent ion (nZVI) from water. Norwegian University of Life Science, Ås, 2014.

34.   Rohwer, F.; Edwards, R., The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.* **2002**, *184* (16), 4529-4535.

35.   Ruokoranta, T. M.; Grahn, A. M.; Ravantti, J. J.; Poranen, M. M.; Bamford, D. H., Complete genome sequence of the broad host range single-stranded RNA phage PRR1 places it in the Levivirus genus with characteristics shared with Alloleviviruses. *J. Virol.* **2006**, *80* (18), 9326-9330.

36.   Salton, M. R.; Kim, K.-S.; Baron, S., *Medical Microbiology*. University of Texas Medical Branch at Galveston: Galveston, TX, **1996**.

37.   Tidona, C.; Darai, G., *The Springer index of viruses*. Springer Science & Business Media: **2011**.

38.   Vaucher, R. d. A.; Dezen, D.; Simonetti, A. B.; Spilki, F. R.; Roehe, P. M., Phylogenetic characterization of bovine parainfluenza 3 from contaminated cell cultures and field isolates from Brazil. *Braz. J. Microbiol.* **2011**, *42* (4), 1440-1444.

39.   Vinnerås, B.; Nordin, A.; Niwagaba, C.; Nyberg, K., Inactivation of bacteria and viruses in human urine depending on temperature and dilution rate. *Water Res.* **2008**, *42* (15), 4067-4074.

40.   Wikipedia, Bacteriophage. **2016**, https://en.wikipedia.org/wiki/Bacteriophage (2016-5-18).

41.   Wikipedia, Transmissible gastroenteritis coronavirus. **2016**, https://en.wikipedia.org/wiki/Transmissible_gastroenteritis_coronavirus (2016-5-18).

42.   Wikipedia, Bovine virus diarrhea. **2016**, https://en.wikipedia.org/wiki/Bovine_virus_diarrhea (2016-5-18).

43.   Wikipedia, Mouse hepatitis virus. **2015**, https://en.wikipedia.org/wiki/Mouse_hepatitis_virus (2016-5-18).

44.   Wikipedia, T3 phage. **2016**, https://en.wikipedia.org/wiki/T3_phage (2016-5-18).

45.   Xie, Z.; Zhang, M.; Xie, L.; Luo, S.; Liu, J.; Deng, X.; Xie, Z.; Pang, Y.; Fan, Q., Identification of a triple-reassortant H1N1 swine influenza virus in a southern China pig. *Genome Announcements* **2014**, *2* (2), e00229-14.

46.   Ye, Y.; Ellenberg, R. M.; Graham, K. E.; Wigginton, K. R., Survivability, Partitioning, and Recovery of Enveloped Viruses in Untreated Municipal Wastewater. *Environ. Sci. Technol.* **2016**, *50* (10), 5077-85.

47.   Zhang, C.; Li, W.; Liu, W.; Zou, L.; Yan, C.; Lu, K.; Ren, H., T4-like phage Bp7, a potential antimicrobial agent for controlling drug-resistant Escherichia coli in chickens. *Appl. Environ. Microbiol.* **2013**, *79* (18), 5559-5565.

Section S2. Exploratory Data Analysis, whether predictors meet tests of independence and response variable is sufficiently a Normal distribution.

Table S1. Data summary of predictor and response variable attributes in Model 1. 464 observations.

Predictors	% value = 0	% value = 1	% value = 2	% value = 3
Contam. Cat.	NA			
DNA virus (y)			NA	NA
Env. virus (y)			NA	NA
F-Oral (y)			NA	NA

	Mean	Std dev.	Min	Max	% of obs ≤ 1std dev from mean	% of obs ≤ 2 std devs from mean
Predictor						
Temperature	22.23	11.67	0	55	62.7%	96.5%
Response var.						
Sqrt (log (t90sec))	2.33	0.2696	1.248	2.801	25.9%	95.3%

Note: NA means that suggested value is out of possible range.

==

Pearson correlation statistics (correl. command in STATA, see http://www.stata.com/manuals13/rcorrelate.pdf), for the input variables used in Model 1, are shown in Table S2. Note low evidence of correlation, which is also plausible evidence of independence: the absolute value of Pearson statistics are almost all below 0.30. The only exception is that viruses with primary transmission method = faecal-oral tend not to be enveloped (correl. statistic = -0.7077).

Table S2. Pearson correlation statistics between predictors in Model 2.

Predictor	Contamination Category (1, 2 or 3)	DNA Virus (y= 1)	Enveloped Virus (y= 1)	Faecal-Oral is Primary Transmission Path (y=1)	Temperature (C)
Contam. Cat.	1.00				
DNA virus (y)	0.05	1.00			
Env. virus (y)	-0.02	-0.28	1.00		
F-Oral (y)	< 0.00	0.14	**-0.71**	1.00	
Temperature	0.05	0.03	-0.08	-0.02	1.00

Figure S1 Normality of the dependent variable. The frequencies (count of observations in equal width ranges) of values for response variable [square root of logarithmic transformation of inactivation time in seconds, or sqrt ($\log_{10}$ [t90sec])] is show in histogram plot below.
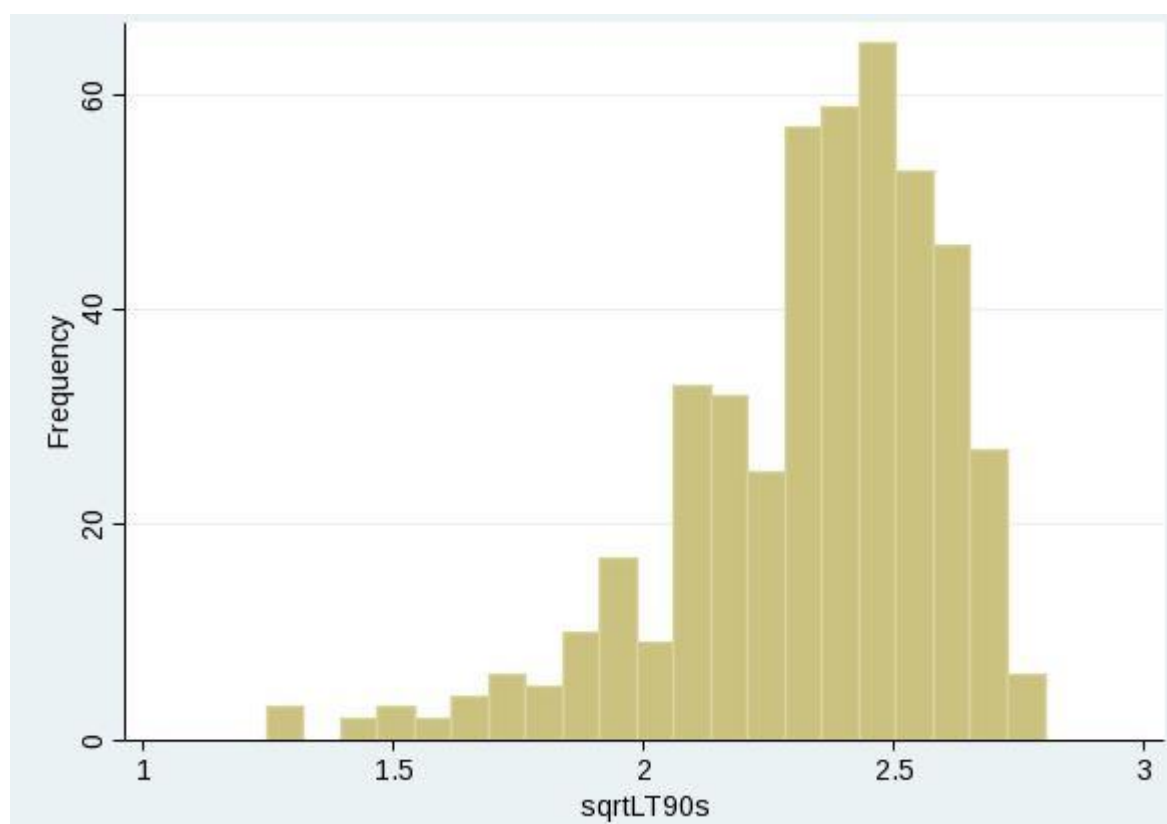
Note: sqrtLT90s = response variable, square root [$\log_{10}$ (T90 in seconds)].

The distribution is visibly left skewed; this is due to the missing right hand tail (n=73/464, or 15.7% of values were known under-estimates of the true T90, due to censored data, arising from upper limits of observation periods). For response variable (see Table S2), mean = 2.33, std dev = 0.27. 25.9% of response observations are within 1 std dev of mean (which is well below the target threshold for assuming a "normal" distribution, of 68%; [Meinrath, Günther, and Petra Schneider. Quality assurance for chemistry and environmental science: metrology from pH measurement to nuclear waste disposal. Springer Science & Business Media, 2007]), but 95.3% of observations are within 2 std. deviations (meets the threshold target of $\geq 95\%$). To treat the true underlying distribution of real inactivation times as relatively Gaussian is defensible.

1

Section S3. Best fit model that incorporates pH (in form of variable pHdiff7) as a predictor.

The model construction and output from Stata appear below.

Table S3. Best fit model that incorporates pH (in form of variable pHdiff7) as a predictor.

```
. xi: cnreg sqrtlt90s faecal_oral_y envelope_y dna_y temp i.contamcat phdiff7, censored(censored)
i.contamcat        _Icontamcat_1-3    (naturally coded; _Icontamcat_1 omitted)

Censored-normal regression                    Number of obs    =        323
                                              LR chi2(7)       =     228.93
                                              Prob > chi2      =     0.0000
_og likelihood =  67.069635                   Pseudo R2        =     2.4151
```

sqrtlt90s	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
faecal_oral_y	.1955415	.0285772	6.84	0.000	.1393159	.251767
envelope_y	-.1081121	.0286013	-3.78	0.000	-.1643851	-.0518391
dna_y	.0129125	.0227695	0.57	0.571	-.0318864	.0577114
temp	-.0085323	.0009073	-9.40	0.000	-.0103173	-.0067472
_Icontamcat_2	-.0463994	.0297125	-1.56	0.119	-.1048586	.0120599
_Icontamcat_3	-.2152351	.0266429	-8.08	0.000	-.267655	-.1628152
phdiff7	.0316144	.0169665	1.86	0.063	-.0017671	.0649959
_cons	2.509265	.0386265	64.96	0.000	2.433268	2.585263
/sigma	.1746579	.0074073			.160084	.1892317

```
    1  left-censored observation
  275     uncensored observations
   47 right-censored observations
```

Variables are:
Faecal_oral_y = faecal oral (or not)
Envelope_y = enveloped virus (or not)
DNA_y = DNA virus (else RNA)
Temp = temperature in C°
Phdiff7 = absolute difference of pH from 7.0
_cons = constant

The model defaults to contamination category 1 (low level of contamination, as defined in Methods in article).

_Icontamcat_2 = adjustment when contamination level = medium
_Icontamcat_3 − adjustment when contamination level = high

The robust standard error was 0.0130835, determined using a clustered sandwich estimator for the standard deviation of model residuals (vce option in Stata, clustering by genus):

Table S4. Robust standard error for Model using phdiff7 as a predictor variable.

```
. mean Residuals, vce(cluster genus)

Mean estimation                    Number of obs   =      275

                    (Std. Err. adjusted for 25 clusters in genus)

                              Robust
                  Mean      Std. Err.     [95% Conf. Interval]

     Residuals  -.013324    .0130835     -.040327     .013679
```

The residual standard error in this Model is superior (lower) than the preferred Model 1 discussed in the article. However, there are many caveats to consider, and why the Model in the main article was preferred:

1. pH in real world matrices often varies over time; our extracted pH values = the average over the observation period. However, the real world average pH in a specific media could be impossible to predict prospectively. Admittedly, in real world environments, another predictor, temperature, will also likely vary. However, temperature fluctuations may be easier to predict and monitor.

2. pH in the above model increases T90 (is supportive of longer virus life) with distance from pH=7.0. This result is counter to expectation, and probably reflects the relatively limited number of pH values in the input data: 57% of records where pH was stated, had pH=6-8 and 42% of records had pH = 8-9.3.

3. The above model may have a lower robust standard error mostly because of omitting specific outliers, and the reduction in number of clusters (from 32 to 25). Also, the model presented in this supplemental file describes a narrower range of conditions and virus species.

Section S4. Impact of Influential Observations.

The leverage statistic h_i and Cook's distance (http://www.columbia.edu/~so33/SusDev/Lecture_5.pdf, http://www.real-statistics.com/multiple-regression/outliers-and-influencers/, http://stat.ethz.ch/~mmarloes/teaching/stat423/handouts/Diagnostics.pdf) were calculated for 389 uncensored observations used to generate Model 1. Any observation exceeding relatively high (1.8 x std devn) for any two of Cook's distance, h_i and residual, was identified as "influential". The eligible 11 observations are detailed below. Note that most have relatively extreme temperatures (≤ 7 or ≥ 50 deg. C); ineligible records (for selection as "influential") with high residuals but much lower leverage/Cook's distance values had temperatures in the range of 10-25°.

Table S5. Eleven most influential observations in Model 1.

Virus	DNA_y	Env_y	FO_y	Temp	pH	Contam'n Category	leverage, h_i	Cook's D (distance)
ADV	1	1	0	50	7.6	High	0.014376	0.05837
ADV	1	1	0	50	7.2	Low	0.014376	0.02732
FMDV	0	0	0	55	8.3	High	0.019174	0.05693
FMDV	0	0	0	55	7.2	Low	0.019174	0.06844
HAdV2	1	0	0	7	7.64	High	0.005832	0.02624
HAdV2	1	0	0	7	7.57	Medium	0.005832	0.04138
HAdV2	1	0	0	7	7.57	Low	0.005832	0.03814
HAdV2	1	0	0	7	7.35	Low	0.005832	0.02959
MS2	0	0	1	10	unclear	Medium	0.004526	0.01864
SIV	0	1	0	55	8.3	High	0.019174	0.05964
SIV	0	1	0	55	7.2	Low	0.019174	0.06213

Notes: ADV = pseudorabies = Aujeszky's disease virus; FMDV = foot-and-mouth disease virus. HAdV2 = Human adenovirus type 2; MS2 = a common phage; SIV = swine influenza.

Regenerating Model 1 without the 11 influential observations, yielded:

Table S6. Model coefficients and statistics with 453/464 input records.

		95% CI for coeff. values		
	Coefficient	Lower bound	Upper bound	p-value
Model constant	2.62413	2.5531	2.6951	< 0.000
Faecal oral transmission pathway (y)	0.59007	0.0061	0.1119	0.029
Enveloped virus (y)	-0.13726	-0.1910	-0.0835	< 0.000
DNA virus (y)	0.03322	-0.0075	0.0740	0.110
Temperature in C°	-0.00910	-0.0107	-0.0075	< 0.000
Low contamination	0	Na	na	
Medium contamination	0.00354	-0.0413	0.0484	0.877
High contamination	-0.11404	-0.1557	-0.0724	< 0.000

Notes: sqrt(T90secs) = square root[$\log_{10}$(T90 in seconds)]. Enveloped virus (y) = 1 when enveloped, else 0. Faecal oral (y) = 1 when faecal oral is primary transmission pathway, else 0. DNA virus (y) = 1 for DNA virus, else 0. Model default is when level of contamination = low, else model adjusts for when contamination is medium or high as indicated.

The 95% confidence intervals for all predictor coefficients are all narrower than those for Model 1 (in main article) but still overlap considerably. The percentage overlap (Model 1 range as denominator) is mostly > 50%, as shown in the table below, values from 34.8% to 92.2%. All of the mean predictor coefficients in Model 1 are within the 95% confidence intervals obtained in the minus-influential observations model as shown above and below, except for the faecal oral transmission variable. Model 1 gives a mean estimate of the FO coefficient = 0.12877, which is outside the 95% CI of the alternative model mean estimate for FO pathway (95% CI ranges from 0.0061 to 0.1119). This difference along with the temperature extremes observed in the most influential observations and the reduced significance (p-value for the FO variable), suggests that the FO-transmission pathway matters less at temperature extremes. This finding links to our hypothesis that more data for observations at relatively extreme temperatures could allow for relating temperature to T90 in a different relationship (not linear), and would probably improve model fit and reduce uncertainty.

Overall, we conclude that none of the 11 influential observations are so influential as to merit complete removal from Model 1.

Table S7. 95% confidence intervals for coefficients in Model 1 (main manuscript) and alternate model without 11 influential observations.

| Variable | Model 1 95% CI | | =ub-lb | Alternate model minus influent. Obs 95% CI | | | =W. Overlapping/ Model 1 Width |
	Lower bound	Upper bound	Model 1 Width	Lower bound	Upper bound	Width overlapping	% Overlap of both confidence intervals
constant	2.4946	2.643	0.1485	2.5531	2.6951	0.08996	60.5%
Temp	-0.0114	-0.0080	0.0033	-0.0107	0.0075	0.00266	80.4%
FO_y	0.0730	0.1845	0.1114	0.0061	0.1119	0.04565	34.8%
Env_y	-0.1509	-0.0392	0.1117	-0.1911	-0.8348	0.1076	60.4%
DNA_y	-0.0287	0.0592	0.0879	-0.0076	0.0740	0.06671	75.9%
Med contam	-0.0447	0.0532	0.0979	-0.0413	0.0484	0.08977	91.6%
High contam	-0.1579	-0.0675	0.0904	-0.1557	-0.0724	0.08331	92.2%

Notes: Temp in C°, FO_y = virus is primarily faecally-orally transmitted. Env_y = enveloped virus. DNA_y = DNA not RNA nucleic acid. Other categorical indicators = Medium or high levels of contamination.

Section S5. 95% confidence intervals for T90 of viruses in medium contaminated matrices.

Model 1 uncertainty is high and hence confidence intervals for inactivation rates (T90 values) are quite wide, as these graphs show. Note that the horizontal scale is consistent (from 0 to 44 degrees), but the vertical scale varies hugely. Viruses are grouped (A-H) by combination of categorical traits. T90s are shown for each group as a function of temperature for viruses in matrices with a medium level of contamination. See main article for more details.

*Note*s *for Figure S2*: Dark orange = below mean estimate, light orange = above mean estimate. Data are for Groups A-H as defined in main article, and mid point estimates shown in Figure 3b in main article. The panels in Figure S2 visually demonstrates the wide confidence intervals and high model uncertainty; high uncertainty about model predictions of T90 is also true for matrices with low or high levels of contamination (Figures 3a and 3c in main article).

Figure S2. Below mean (dark orange) and above mean (lighter orange) shares of 95% confidence intervals for T90s in matrices with medium levels of contamination.